Wireless Communications and Mobile Computing

Mobile Intelligence Assisted by Data Analytics and Cognitive Computing 2020

Lead Guest Editor: Yin Zhang
Guest Editors: Huimin Lu and Haider Abbas

Mobile Intelligence Assisted by Data Analytics and Cognitive Computing 2020

Wireless Communications and Mobile Computing

Mobile Intelligence Assisted by Data Analytics and Cognitive Computing 2020

Lead Guest Editor: Yin Zhang
Guest Editors: Huimin Lu and Haider Abbas

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8891598, 12 pages
https://doi.org/10.1155/2020/8891598

Research Article

A Novel Adaptive Directional Interpolation Algorithm for Digital Video Resolution Enhancement

Dong Sun,[1] **Qingqing Xie**,[1] **Teng Li**,[1] **Yixiang Lu**,[1] **De Zhu**,[2] **and Qingwei Gao**[1]

[1]*College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui, China 230601*
[2]*Network and Information Center, Anhui University, Hefei, Anhui, China 230601*

Correspondence should be addressed to Qingwei Gao; qingweigao@ahu.edu.cn

Received 25 June 2020; Accepted 23 November 2020; Published 16 December 2020

Academic Editor: Huimin Lu

In this paper, a novel digital video resolution enhancement algorithm based on adaptive directional interpolation is proposed, where the directionality of the edge structure and the nonlocal self-similarity prior within the current frame as well as its adjacent frames are both considered. First, we establish the regularization equation that conforms to the prior model of a video frame and then take the classic bicubic interpolation result as the initial estimation to iteratively solve the restoration equation, in which the edge structures and contours in low resolution (LR) input are reconstructed to estimate and refine the desired high resolution (HR) output. Experimental results show that the proposed algorithm can effectively enhance the clarity of a video frame, with satisfying subjective visual quality and PSNR value.

1. Introduction

Videos and images are the main sources of information for humans. According to statistics, more than 80% of the information we receive from the outside world comes from vision. With the development of digital mobile communication and computer technology, various novel applications such as distance education, video on demand, telemedicine, and multiperson online video conference have appeared, promoting the revolution of productivity and social progress. In the meantime, the image quality of digital video has also been desired higher and higher, where the clarity index comes from standard definition to high definition (HD) and ultra-high definition, as well as the corresponding resolution index also comes from 480p to 720p, 1080p, and 2160p (4K). On the one hand, these improvements in clarity and resolution can meet the increasing demand of end users and provide better image quality; on the other hand, while high-resolution video provides more details in content, it also adds burdens to the entire production and consumption ecosystem: more expensive capture and storage devices on the image acquisition side, additional computing resource requirement for video editing on the media creation side, and more data transmission pressure on the communication network side. All these above have become important factors that restrict further improvement of video clarity and quality. In order to solve this problem, a common way is to use an image postprocessing procedure where the LR input frame is interpolated by a superresolution method [1–7], leading to a resolution-enhanced HR one. This software-based technique does not change the existing image acquisition and data transmission systems and thus is of great value in fields of videotelephony, virtual reality, augmented reality, and HD video games.

Natural images are highly structured, which reflects the strong time-spatial redundancy and self-similarity underlying pixels and performs a key role in solving inverse problems such as image denoising, deblurring, inpainting, and superresolution. By considering the fact that the human visual system is sensitive to the image edge structure [7–11], a novel digital video resolution enhancement algorithm via adaptive directional filtering is proposed in

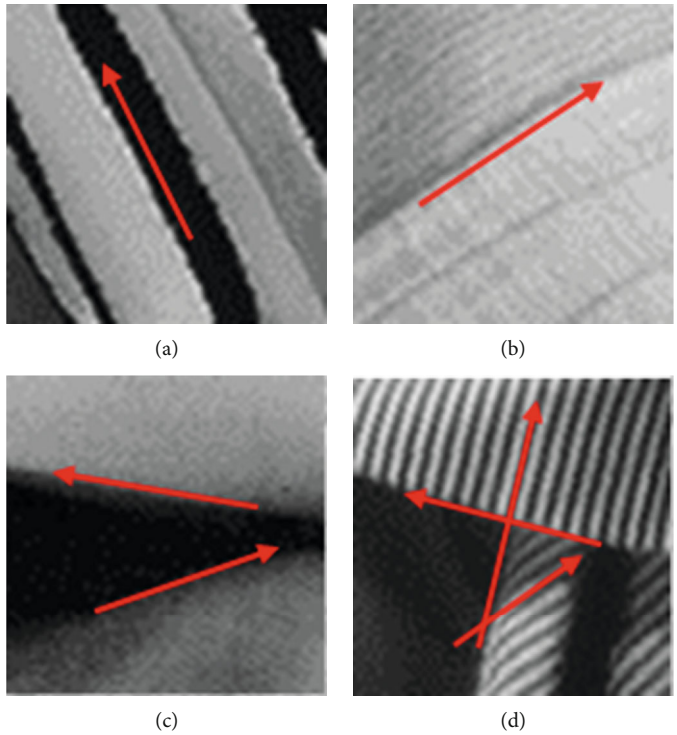


Figure 1: Directional regularity in natural images: (a) an image patch contains one main direction (120 degrees); (b) an image patch contains one main direction (40 degrees); (c) an image patch contains two main directions (30 degrees and 170 degrees); (d) an image patch contains three main directions (40 degrees, 80 degrees, and 165 degrees).

this paper, in which the characteristics of the edge contour and the nonlocal self-similarity within current frame as well as the corresponding adjacent frames are both considered. We first establish the regularization equation that conforms to the prior model of a video frame and then take the classic bicubic interpolation result as the initial estimation to iteratively solve the restoration equation, where the edge structures and contours in LR input are reconstructed to estimate and refine the desired HR output.

The rest of the sections are organized as follows. In Section 2, we introduce the core idea of the proposed adaptive directional interpolation scheme for estimating the missing details of the LR image and then use the nonlocal self-similarity prior to further improve the interpolation performance. The details of the video resolution enhancement algorithm are provided in Section 3. Section 4 presents the experimental validations of the proposed algorithm and comparison with the classic bicubic interpolation method; conclusions are drawn in Section 5.

2. The Core Idea

Directional regularity has widely existed in textures, edges, and contours of natural images (shown as in Figure 1). Denote vector $f_i \in \mathbb{R}^{n^2}$ as the image patch centered around the ith pixel and with sizes $n \times n$, and $L_\theta \in \mathbb{R}^{n^2 \times n^2}$ as the filter matrix corresponding to the directional filter with angle θ (in this paper, the directional controllable steerable filter [12] is used). Obviously, the filtered vector $L_\theta f_i$ is the sparsest (namely, $L_\theta f_i$ is approximate to zero) when θ is parallel with the main direction of f_i. Generally, an image patch may include more than one main direction due to its complexity (examples are shown in Figures 1(c) and 1(d)); we can search for these direction angles using the following algorithm:

In our previous works [3, 13], we have shown the details to construct a blurring matrix from its corresponding linear degradation operator (as well as the downsampling matrix $\mathbf{H}$). Here, we simply present the steps to construct the directional filter matrix $\mathbf{L}$ from a 2-D filter kernel $\mathbf{B}$, as follows:

(i) Let $\mathbf{L}$ be a $n^2 \times n^2$ zero matrix;

(ii) For each pixel of the filtered image patch $\mathbf{d} = \mathbf{L}\mathbf{f}_i$:

(a) Compute the 2-D coordinate (r, c) of pixel $d[i]$ from its 1-D index i;

(b) For each element $B[u][v]$ of filter kernel $\mathbf{B}$, set the element $L[i][(c - v - 1)n^2 + r - u] = B[u][v]$.

The structure of filter matrix $\mathbf{L}$ is presented in Figure 2.

Figure 3 shows the main direction searching results of test images *barbara* and *butterfly* using the algorithm above.

Denote $\mathbf{y}_i = \mathbf{H}\mathbf{f}_i$ as the LR image patch, where $\mathbf{H} \in \mathbb{R}^{m^2 \times n^2}$ is the downsampling matrix [3]. When the downsampling factor D is an integer, we have $m = n/D$, and the corresponding LR input can be represented as $y(h, v) = f(h/D, v/D)$. With the constraint of the directional regularity posed above,

Main direction searching
Partition $\mathbf{f}$ into overlapping patches $\{f_1, f_2, \cdots\}$, and for each patch, do the following steps:

- **Initialization:** Set main direction angle set $\mathbf{S} = \varnothing$, candidate angle set $\mathbf{\Theta} = \{\theta_1, \theta_2, \cdots \theta_K\}$, the largest number of direction angles P. Set start point $\mathbf{d} = \mathbf{f}_i$.
- **Main loop** (repeat P times):
  - Calculate the filtering result $\mathbf{L}_{\theta 1}\mathbf{d}, \mathbf{L}_{\theta 2}\mathbf{d}, \cdots \mathbf{L}_{\theta K}\mathbf{d}$;
  - Find the best angle $\theta_{opt} = \text{argmin}_{\theta_j} \|\mathbf{L}_{\theta j}\mathbf{d}\|_1$;
  - Update $\mathbf{S} \leftarrow \mathbf{S} \cup \{\theta_{opt}\}$, $\mathbf{\Theta} \leftarrow \mathbf{\Theta}/\{\theta_{opt}\}$ and $\mathbf{d} \leftarrow \mathbf{L}_{\theta_{opt}}\mathbf{d}$ for the next iteration.
- **Output:** The main direction angle set $\mathbf{S}$ of the ith image patch $\mathbf{f}_i$.

Algorithm 1:

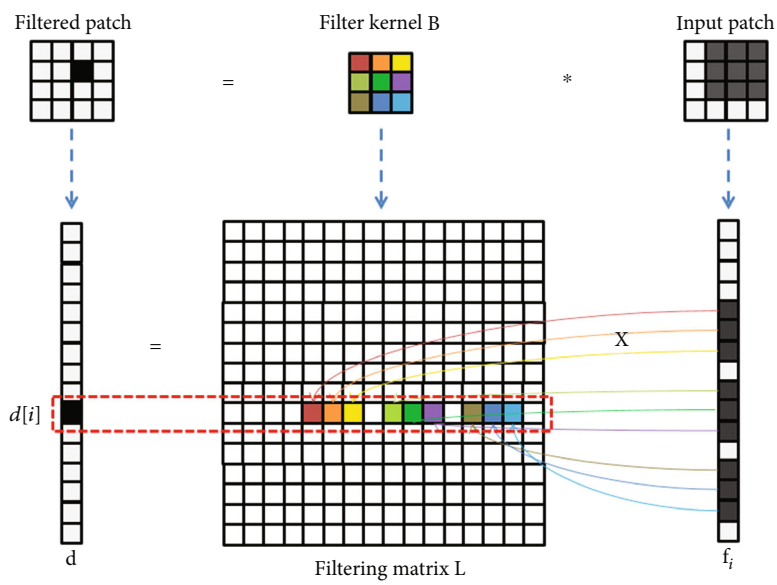
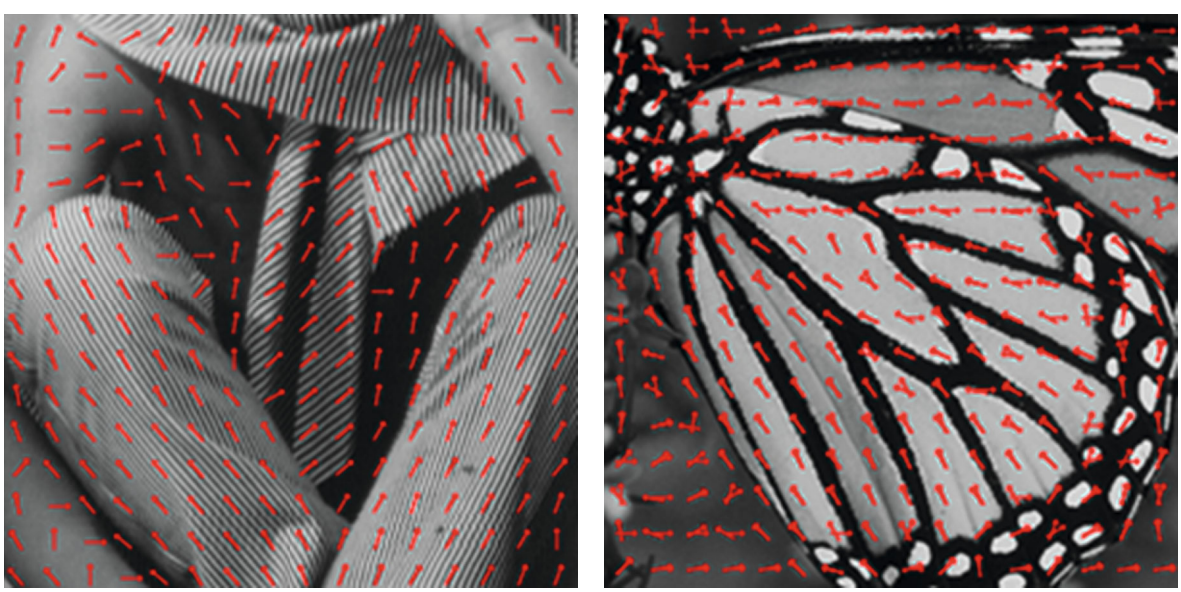


Figure 2: Structure of the directional filter matrix.



(a) (b)

Figure 3: Main direction searching results: (a) *barbara* ($P = 1$); (b) *butterfly* ($P = 2$. Directions with similar degrees are merged).

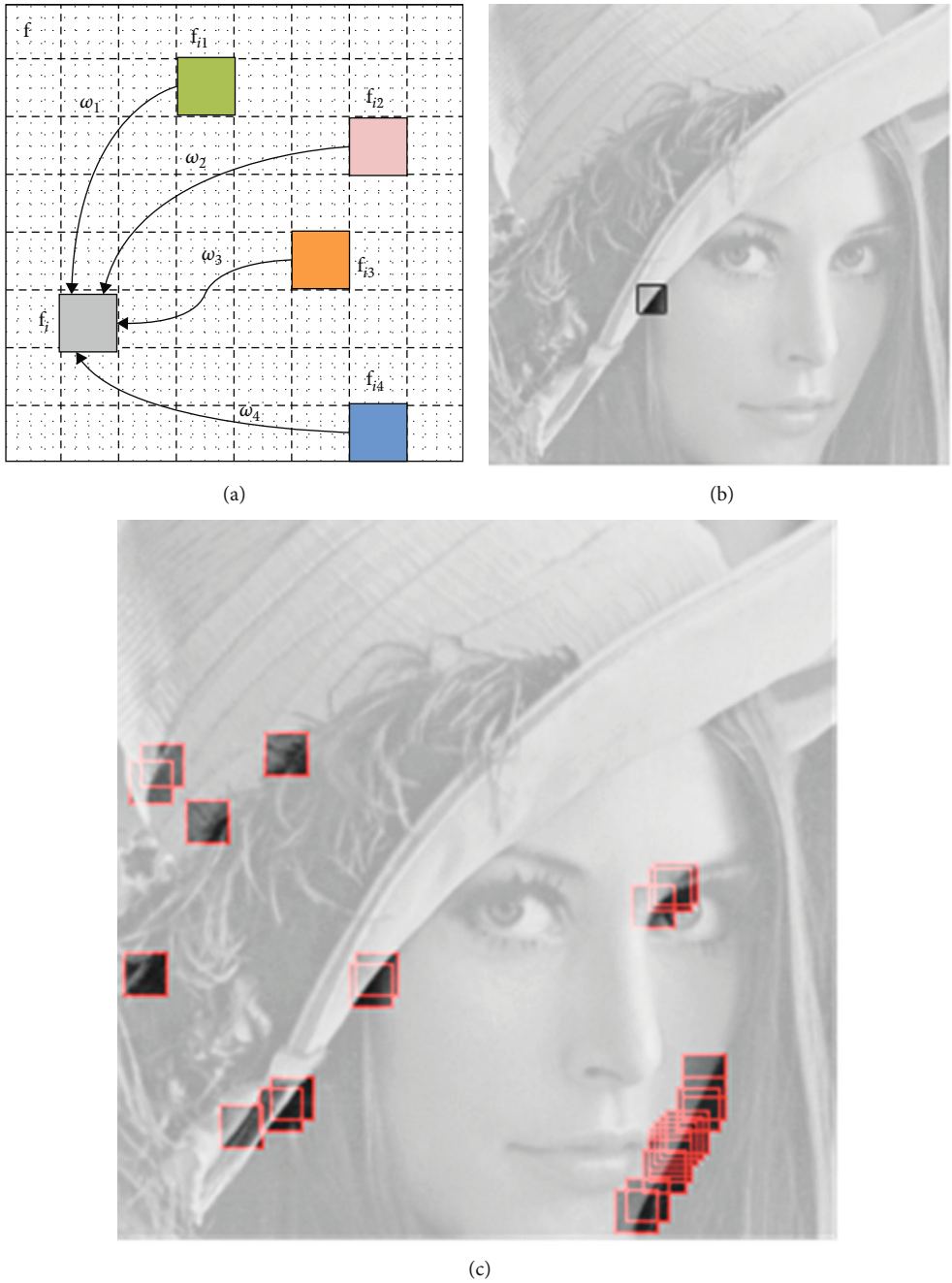


Figure 4: The NAR image model. As an example of (a), one image patch in (b) can be linearly represented by several nonlocal neighbors in (c).

the following interpolation equation can be used to estimate the original HR patch $\mathbf{f}_i$ that

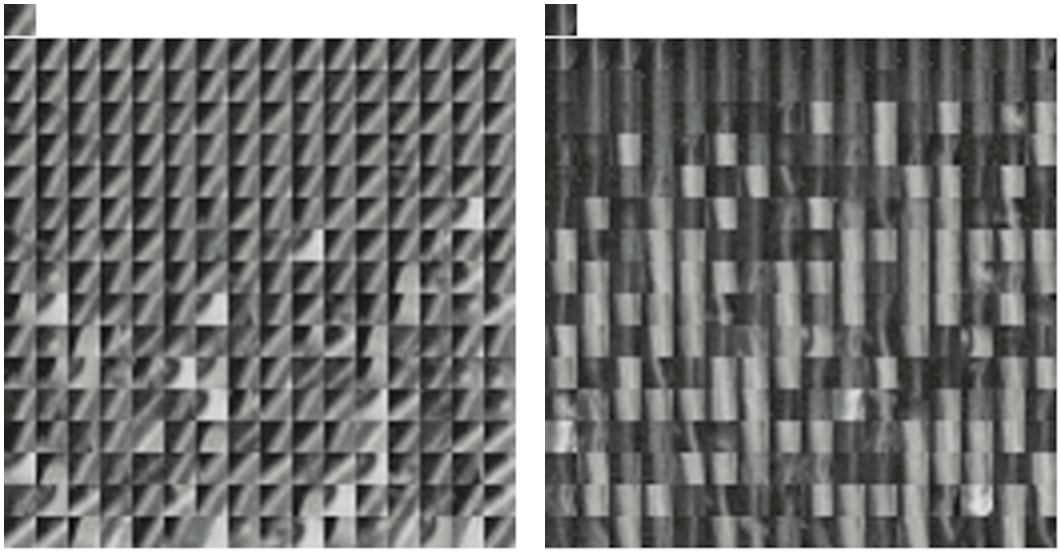
$$\hat{\mathbf{f}}_i = \operatorname{argmin}_{\mathbf{f}_i} \|\mathbf{y}_i - \mathbf{H}\mathbf{f}_i\|_2^2 + \lambda \|\mathbf{L}_i \mathbf{f}_i\|_2^2, \tag{1}$$

where λ is the regularization parameter and $\mathbf{L}_i = \prod_{p=1}^{P} \mathbf{L}_{\theta p}$ is the adaptive directional filter matrix. This equation posed above has the well-known closed-form solution

$$\hat{\mathbf{f}}_i = \left(\mathbf{H}^T\mathbf{H} + \lambda \mathbf{L}_i^T \mathbf{L}_i\right)^{-1} \mathbf{H}^T \mathbf{y}_i. \tag{2}$$

It is easy to know from the structure of the downsampling matrix $\mathbf{H}$ that $\mathbf{H}^T\mathbf{H}$ is diagonal. For the downsampling factor $D = 2$, we have

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} \mathbf{E}_n & & & & \\ & \mathbf{0}_n & & & \\ & & \cdots & & \\ & & & \mathbf{E}_n & \\ & & & & \mathbf{0}_n \end{bmatrix} \in \mathbb{R}^{n^2} \times n^2, \tag{3}$$



(a) (b)

Figure 5: Image patch (top-left, $n = 8 \times 8$) and its adaptive dictionary (bottom, $M = 256$).

where $\mathbf{E}_n = \text{diag}\,(1, 0, 1, 0, \cdots, 1, 0) \in \mathbb{R}^{n\times n}$ and $\mathbf{0}_n \in \mathbb{R}^{n\times n}$ is a zero matrix. Plugging the SVD decomposition $\mathbf{L}_i^T\mathbf{L}_i = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T = \mathbf{U}\,\text{diag}\,(\sigma_1, \sigma_2, \cdots, \sigma_{n^2})\mathbf{U}^T$ into expression (2), this leads to

$$\begin{aligned}\mathbf{f}_{\hat{i}} &= \mathbf{U}\left(\mathbf{H}^T\mathbf{H} + \boldsymbol{\Lambda}\right) - 1\mathbf{U}^T\mathbf{H}^T\mathbf{y}_i \\ &= \begin{bmatrix} 1+\theta_1 & & & & \\ & \theta_2 & & & \\ & & 1+\theta_3 & & \\ & & & \cdots & \\ & & & & \theta_{n^2} \end{bmatrix}^1 \mathbf{U}^T\mathbf{H}^T, \mathbf{y}_i.\end{aligned} \tag{4}$$

Recall that $\mathbf{L}_i\mathbf{f}_i \approx \mathbf{0}$, and thus, $\mathbf{L}_i$ is approximately singular, implying that one or more singular values of $\mathbf{L}_i^T\mathbf{L}_i$ are close to zero, and therefore, the inverse of the restoration kernel $\mathbf{H}^T\mathbf{H} + \lambda\mathbf{L}_i^T\mathbf{L}_i$ is ill-posed that can not be well handled. To solve this problem, we explore the self-similarity prior widely existing in natural images to further improve the interpolation performance. In this paper, the nonlocal autoregressive (NAR) model of images [14] is used to add additional constraint to the restoration kernel and reduce the degree of freedom of desired unknown pixels; this will help to yield a more stable result.

According to our previous works [15–17], we show that each patch in an image can be approximatively represented as a linear combination of M nonlocal neighbors at different locations (shown as in Figure 4) that

$$\mathbf{f}_i \approx \sum_{j=1}^{M} \omega_j \mathbf{f}_{ij} = \mathbf{F}_i\omega_i. \tag{5}$$

The neighbor set $\mathbf{F}_i = [\mathbf{f}_{i1}, \mathbf{f}_{i2}, \cdots, \mathbf{f}_{iM}] \in \mathbb{R}^{n\times M}$ consists of M nonlocal patches around $\mathbf{f}_i$, which can be seen as an adaptive local dictionary that refers to the target vector $\mathbf{f}_i$, and the corresponding representation coefficient ω_i can be easily computed by ridge regression

$$\omega_i = \left(\mathbf{F}_i^T\mathbf{F}_i + \gamma n\mathbf{I}\right)^{-1}\mathbf{F}_i^T\mathbf{f}_i, \tag{6}$$

where the parameter γ is set manually to lead to the best results. Moreover, we have also proved in [15, 17] that ω_i is sparse when the atoms of $\mathbf{F}_i$ are similar to $\mathbf{f}_i$ in terms of normalized inner products. Considering that sparsity is very powerful that is broadly used in solving various inverse problems and has shown the ability to handle the image superresolution task [3, 6, 14, 15, 18], we here propose the following algorithm (Algorithm 2) to construct the adaptive dictionary $\mathbf{F}_i$:

Figure 5 shows the dictionary construction results of two patches of test images *lena* using the algorithm above. For video sequence, the above algorithm is also adapted to construct a dictionary for image patch of frames. At this time, each atom of $\mathbf{F}_i$ comes from those nonlocal neighbors belonging to the current frame and its adjacent Q frames, shown as in Figure 6. Considering that video scene changes smoothly for most time, the differences between neighbor frames are small; this means it will be easier to find more similar candidate patches and thus finally leads to a sparser/better representation coefficient ω_i, which helps in improving the interpolation performance further.

Replacing the constraint posed in (5) by an equivalent penalty and adding it to Equation (1), we obtain

$$\mathbf{f}_{\hat{i}} = \text{argmin}_{\mathbf{f}_i}\|\mathbf{y}_i - \mathbf{H}\mathbf{f}_i\|_2^2 + \lambda\|\mathbf{L}_i\mathbf{f}_i\|_2^2 + \mu\|\mathbf{f}_i - \mathbf{F}_i\omega_i\|_2^2. \tag{7}$$

Combining this equation with Equation (6), we get the desired HR patch estimator

Adaptive dictionary construction
Partition $\mathbf{f}$ into overlapping patches $\{\mathbf{f}_1, \mathbf{f}_2, \cdots\}$, and for each patch, do the following steps:

- **Initialization:** Set nonlocal neighbor number M and search window size W.
- **Dictionary construction:**
  - Sweep over all possible patches $\mathbf{f}_{i1}, \mathbf{f}_{i2}, \cdots$ over the searching window centered around $\mathbf{f}_i$, and compute the normalized candidate atom set $\mathbf{G}_i = [(\mathbf{f}_{i1}/\|\mathbf{f}_{i1}\|), (\mathbf{f}_{i2}/\|\mathbf{f}_{i2}\|), \cdots]$;
  - Compute the normalized inner product vector $\mathbf{r} = \mathbf{G}_i^T \mathbf{f}_i$;
  - Select the atoms with the largest M values in $|\mathbf{r}|$ to construct dictionary $\mathbf{F}_i$.
- **Output:** The adaptive dictionary $\mathbf{F}_i$ of the ith image patch $\mathbf{f}_i$.

Algorithm 2:

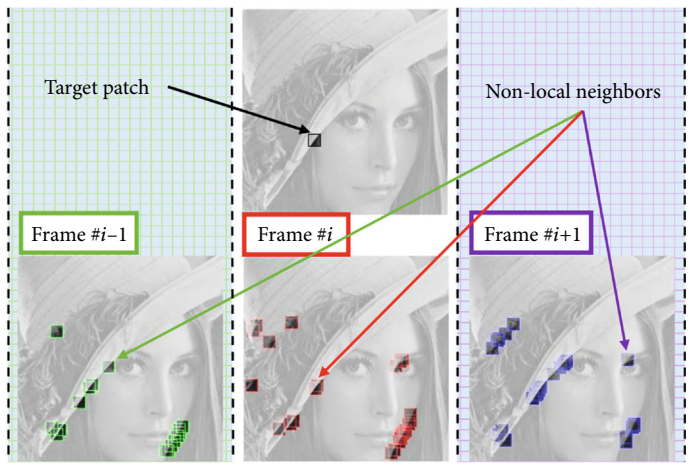


Figure 6: Dictionary construction. As an illustration, for a target patch in the ith frame of a video sequence, the corresponding adaptive dictionary is composed of nonlocal neighbors scattered over frames i, $i-1$, and $i+1$ (take $Q = 1$ for example).

Resolution enhancement algorithm
For each LR frame $\mathbf{y}$ of the input digital video sequence, do the following steps:

- **Initialization:** Set $\mathbf{f}$ the bicubic interpolation of $\mathbf{y}$.
- **Main loop** (repeat C times):
  - Use Algorithm 1 to search the main direction for each patch of $\tilde{\mathbf{f}}$, calculate the corresponding adaptive directional filter matrix $\mathbf{L}_i$;
  - Use Algorithm 2 to construct the adaptive dictionary $\mathbf{F}_i$;
  - Taking $\tilde{\mathbf{f}}_i$ as an initial estimation of the desired HR output $\mathbf{f}_i$, use Equation (8) to compute the resolution enhancement result $\hat{\mathbf{f}}_i$;
  - Update $\tilde{\mathbf{f}} \leftarrow \hat{\mathbf{f}}$ for the next iteration when all image patches have been restored.
- **Output:** The resolution enhanced output $\hat{\mathbf{f}}$.

Algorithm 3:

$$\hat{\mathbf{f}}_i = \left(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{L}_i^T\mathbf{L}_i + \mu\mathbf{I}\right)^{-1}\left(\gamma\mathbf{F}_i\left(\mathbf{F}_i^T\mathbf{F}_i + \gamma n\mathbf{I}\right)^{-1}\mathbf{F}_i^T\mathbf{f}_i + \mathbf{H}^T\mathbf{y}_i\right). \tag{8}$$

Contrast the expression above with formula (2), we can see that the restoration kernel is full rank now, while keeping the advantage of diagonal, leading to a cheap computation of matrix inversion.

3. Video Resolution Enhancement Algorithm

To sum up, we use the interpolation algorithm (Algorithm 3) listed below for digital video resolution enhancement:

A graphic demonstration of this algorithm is displayed in Figure 7.

In each interpolation loop, the time consumption $T_{\text{loop}}(N)$ mainly consists of three parts, including the main direction searching $T_m(N)$, the adaptive dictionary constructing

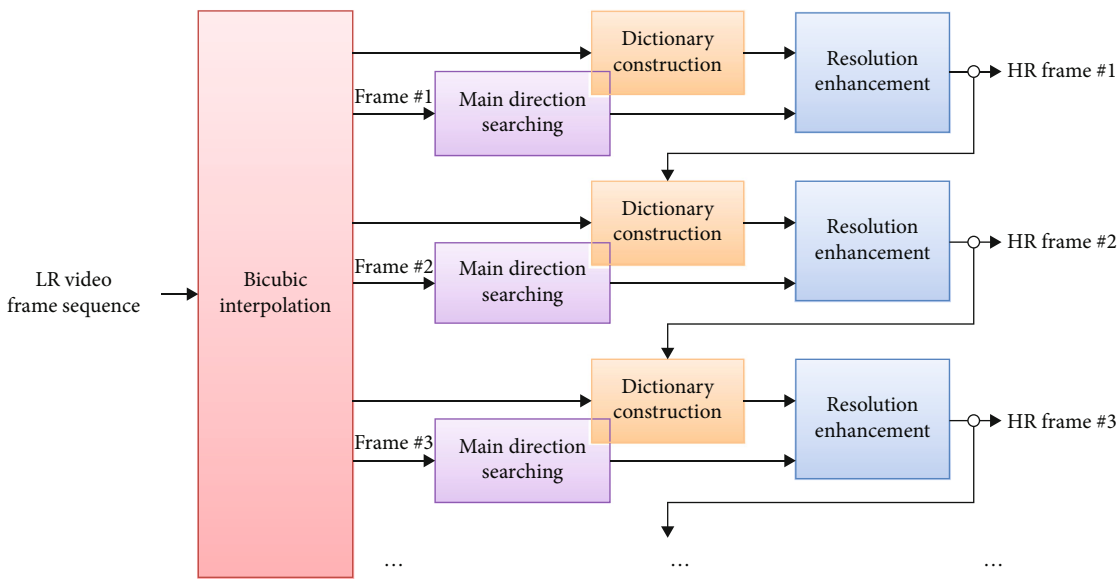


Figure 7: The flowchart of the proposed resolution enhancement algorithm for a video sequence.

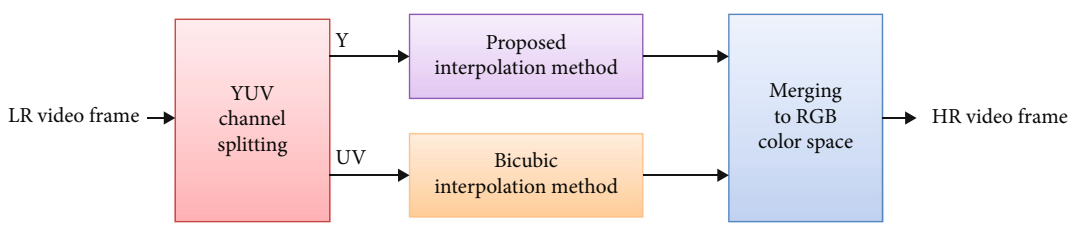


Figure 8: The flowchart of the proposed resolution enhancement algorithm for a color video sequence.

$T_a(N)$, and the HR output estimating $T_e(N)$, where N denotes the size of the LR input frame. That is

$$T_{\text{loop}}(N) = T_m(N) + T_a(N) + T_e(N). \tag{9}$$

For the first term $T_m(N)$, we know from Algorithm 1 that searching each direction for every target patch needs K filtering operations. Considering the fact that filtering a fixed-size image patch with size $n \times n$ can surely be done in constant time t_1, therefore

$$T_m(N) = (N - n + 1)^2 \cdot (P \cdot K \cdot t_1) \sim O(N^2). \tag{10}$$

For the second term $T_a(N)$, we need to sweep over W^2 candidate patches around each target LR patch for searching atoms. Similarly, since the normalization and inner product computing can also be finished in constant time t_2, thus

$$\begin{aligned} T_a(N) = (N - n + 1)^2 \cdot \left(W^2 \cdot t_2 + T_{\text{top}}(M)\right) \sim O(N^2) \\ + O(N^2) \cdot T_{\text{top}}(M). \end{aligned} \tag{11}$$

In the above expression, $T_{\text{top}}(M)$ represents the time consumption of selecting the top M largest elements from vector$|\mathbf{r}| = |\mathbf{G}_i^T \mathbf{f}_i| \in \mathbb{R}^{W^2}$, where this task can be simply implemented by a fast ordering algorithm with time complexity $O(W^2 \log(W))$, and this leads to

$$T_a(N) \sim O(N^2) + O(N^2) \cdot O\left(W^2 \log(W)\right)) \sim O(N^2). \tag{12}$$

For the last term $T_e(N)$, the time consumption is mainly determined by the computation of the inverse matrices $(\mathbf{H}^T\mathbf{H} + \lambda \mathbf{L}_i^T \mathbf{L}_i + \mu \mathbf{I})^{-1}$ and $(\mathbf{F}_i^T \mathbf{F}_i + \gamma n \mathbf{I})^{-1}$. For the reason that the size of $\mathbf{H}$, $\mathbf{L}_i$, $\mathbf{F}_i$, and $\mathbf{I}$ are fixed and indifferent to N, thus these operations can also be done in constant time t_3. We have

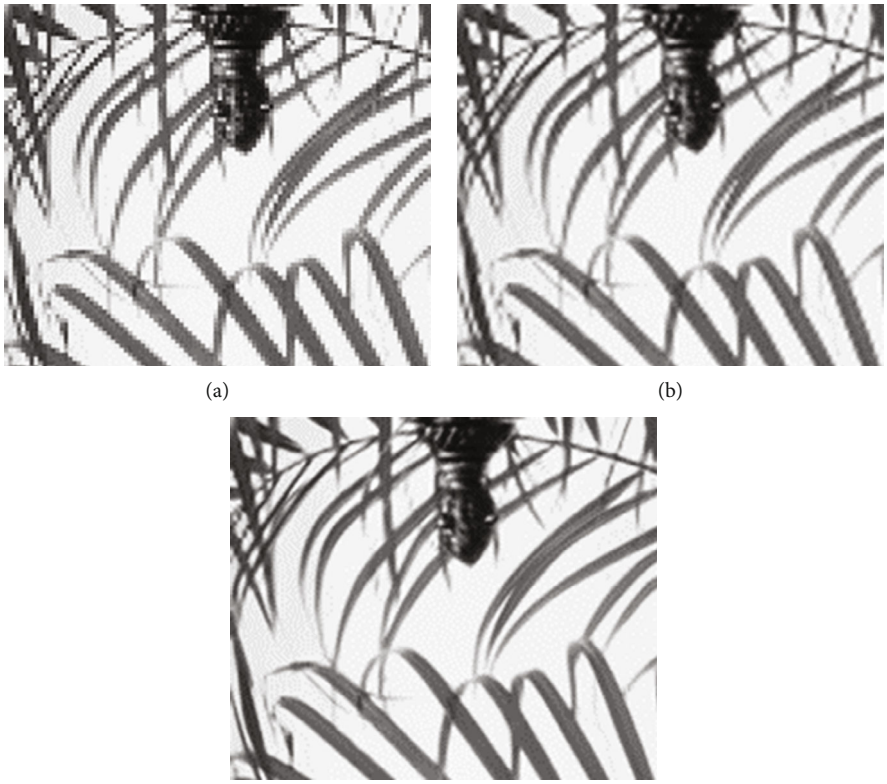
$$T_e(N) = (N - n + 1)^2 \cdot t_3 \sim O(N^2). \tag{13}$$

Plugging Equations (10), (12), and (13) into (9), we obtain

$$T_{\text{loop}}(N) \sim O(N^2) + O(N^2) + O(N^2) \sim O(N^2). \tag{14}$$

The equation above means that the computational complexity of our proposed interpolation algorithm is proportional to the pixel number (N^2) of the LR input frame.

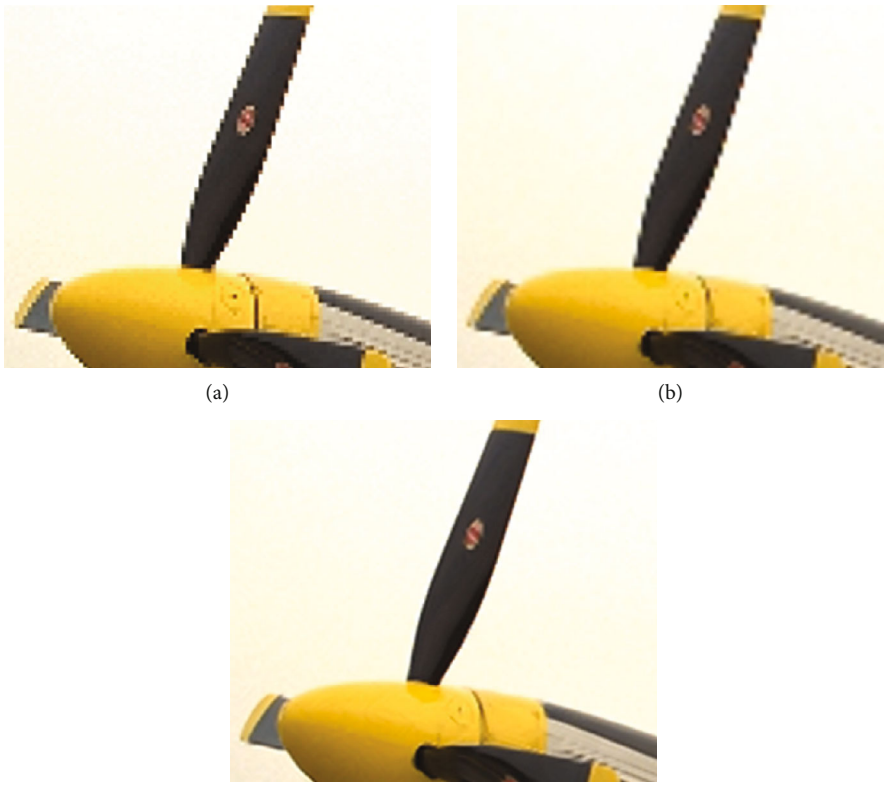
For color video sequence interpolation, the YUV color model can be considered: we start by splitting the input color frame into luminance channel and chrominance channel and then enhance each channel using the proposed algorithm and classic bicubic interpolation, respectively. The final



(a) (b)

(c)

Figure 9: LR image *leaves* and the 2×2 interpolation results: (a) LR image; (b) bicubic (PSNR = 26.64); (c) proposed (PSNR = 29.23).



(a) (b)

(c)

Figure 10: LR image *plane* and the 2×2 interpolation results: (a) LR image; (b) bicubic (PSNR = 29.59); (c) proposed (PSNR = 32.78).

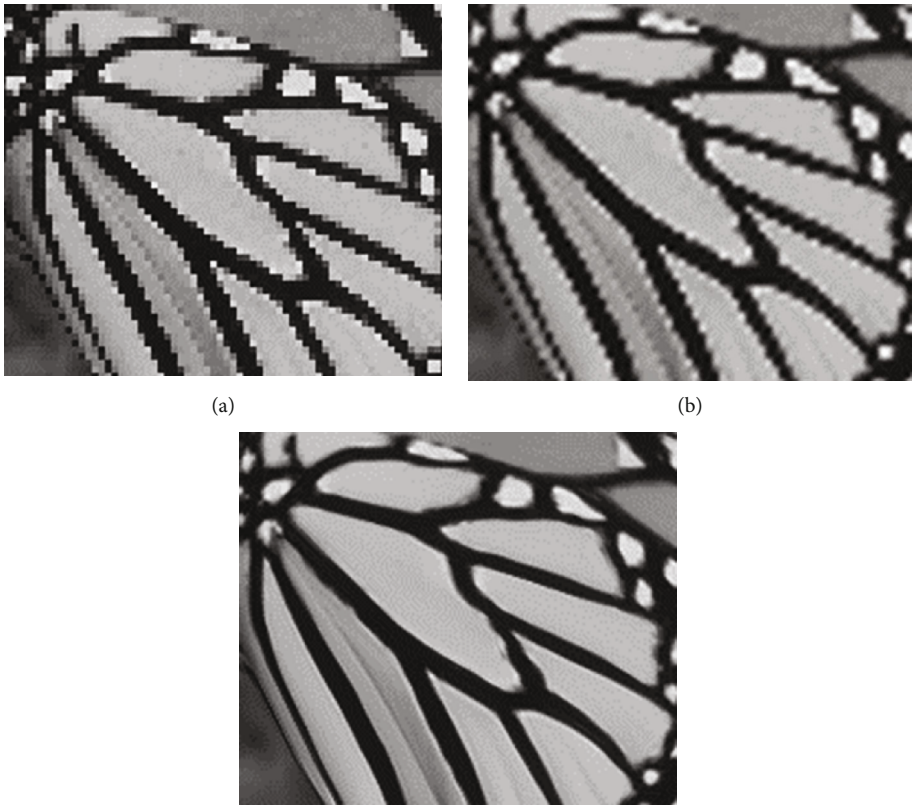


Figure 11: LR image *butterfly* and the 3×3 interpolation results: (a) LR image; (b) bicubic (PSNR = 21.77); (c) proposed (PSNR = 23.45).

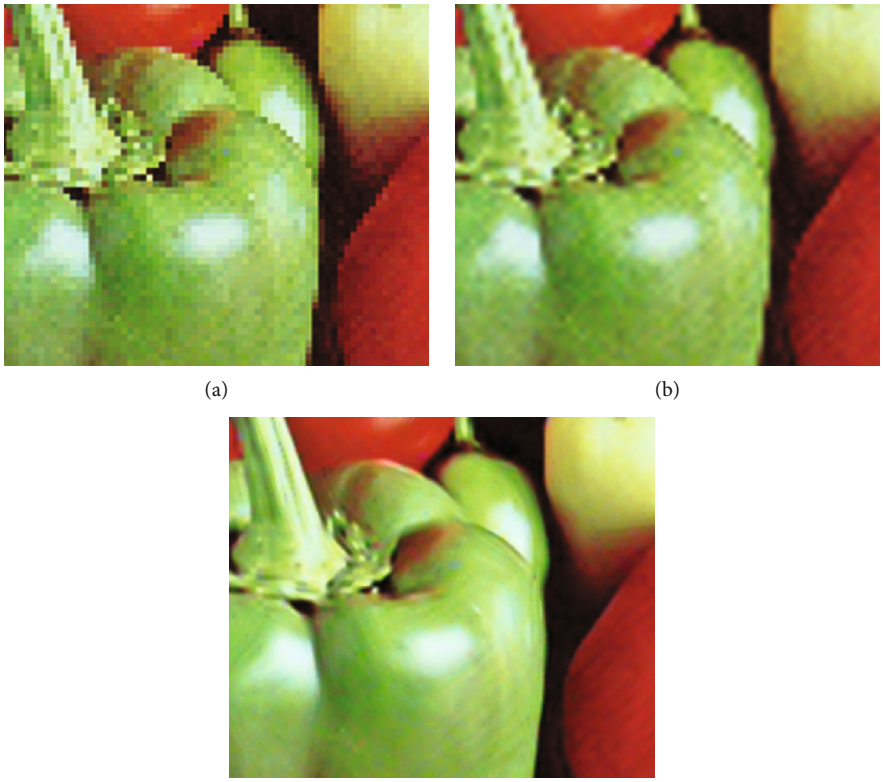


Figure 12: LR image *peppers* and the 3×3 interpolation results: (a) LR image; (b) bicubic (PSNR = 30.75); (c) proposed (PSNR = 32.29).

Figure 13: LR video sequence *foreman* and the 2×2 interpolation results. Row 1: LR frames (#5, #10, #15, #20, and #25). Row 2: the corresponding HR output frames.

Figure 14: LR video sequence *ice* and the 2×2 interpolation results. Row 1: LR frames (#5, #10, #15, #20, and #25). Row 2: the corresponding HR output frames.

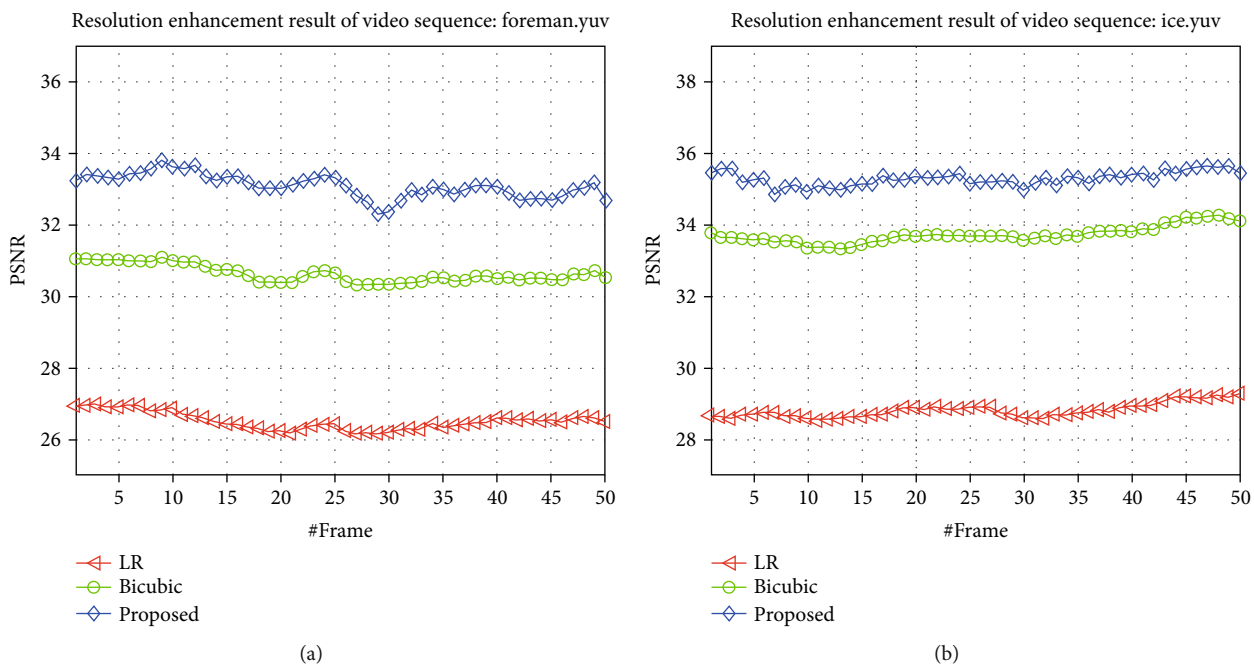


Figure 15: The PSNR evaluation of the 2×2 resolution enhancement outputs: (a) *foreman* (average PSNR = 32.88); (b) *ice* (average PSNR = 35.22).

resolution enhanced frame can be obtained by converting these channels back to RGB color space. The diagram is shown in Figure 8.

4. Experimental Results

In this section, several experimental results of the proposed resolution enhancement algorithm are reported to show the performance and compared with the widely used bicubic interpolation method, in terms of subjective image quality and objective PSNR index. The LR input image/video frame is generated by directly decimating the original HR one by a factor of T in each axis and then interpolated back to the original size for performance evaluation. The chosen parameters are as follows: $P = 2$, $n = 8$, $\lambda = 1$, $\gamma = 800$, $\mu = 5$, $M = 12$, $C = 4$, and $W = 20$, the candidate angle set for main direction searching is $\boldsymbol{\Theta} = \{0,10,20, \cdots 170\}$, and the width of the directional controllable steerable filter is 5 (with Gaussian kernel standard deviation $\sigma = 0.7$). According to our tests, performing a 2×2 interpolation for a single frame costs about 2.3 seconds on Intel Core i7 8750H with 6 cores at 3.9 GHz, Windows 64 bit, Matlab 2017b, accelerated by C-MEX interface in typical settings of $N = 512 \times 512$ and $D = 2$. Using a GPU-accelerated architecture (CUDA or OpenCL) may be helpful to reduce computation time extremely, we shall study this in future research.

Figures 9–12 present the resolution enhancement results on test still images *leaves*, *airplane*, *butterfly*, and *peppers*, with factor $D = 2$ and 3. Figures 13 and 14 further show the 2×2 interpolation results of test video sequences *foreman* and *ice*, with reference frame number $Q = 2$. From these figures, we see that the proposed algorithm works very well in reconstructing image contours and fine details, with few noticeable staircase artifacts in tiny structures, when compared to the bicubic interpolation method which produces a large amount of aliasing in edges and textures, and thus, the performance is very poor. Moreover, Figure 15 also gives the objective quality evaluation of *foreman* and *ice* for the first 50 frames. As expected, our method achieves satisfying PSNR values (with about 2 dBs higher than bicubic on average); this is consistent with the subjective visual quality shown above.

5. Conclusion

In this paper, we present an effective algorithm for enhancing digital video/still image resolution based on the directional regularization and nonlocal self-similarity structure, where the missing pixels of an image patch can be estimated from its nonlocal neighbors via an adaptive directional filtering operation. The appeal of this work is its simplicity, with no requirement of solving complex optimization equations, and is easily implemented. Experimental results show that the proposed algorithm can effectively improve the digital video quality in terms of clarity and resolution and thus will be of great value in theory and application.

Data Availability

Please contact the first author (sundong@ahu.edu.cn) to obtain the Matlab demo codes.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62071001), the Anhui Natural Science Foundation of China (Nos. 2008085MF192 and 2008085MF183), the Key Science Project of Anhui Education Department of China (Nos. KJ2018A0012, KJ2019A0023, and KJ2019A0022), and the CERNET Innovation Project of China (Nos. NGII20180612, NGII20180312, and NGII20180624).

References

[1] T. M. Lehmann, C. Gonner, and K. Spitzer, "Survey: interpolation methods in medical image processing," *IEEE Transactions on Medical Imaging*, vol. 18, no. 11, pp. 1049–1075, 1999.

[2] Z. F. Lu and B. J. Zhong, "Image interpolation algorithm based on prediction gradient," *Acta Automatica Sinica*, vol. 44, no. 6, pp. 1072–1085, 2018.

[3] D. Sun, Q. Gao, Y. Lu, L. Zheng, and H. Wang, "A high quality single-image super-resolution algorithm based on linear Bayesian MAP estimation with sparsity prior," *Digital Signal Processing*, vol. 35, no. 12, pp. 45–52, 2014.

[4] S. Mallat and G. Yu, "Super-resolution with sparse mixing estimators," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2889–2900, 2010.

[5] B. D. Jiang, Z. Xie, and L. Wu, "Multiresolution visualization of coastline using controllable fractal interpolation," *Journal of Computer-Aided Design & Computer Graphics*, vol. 29, no. 11, pp. 2015–2022, 2017.

[6] B. Hou, K. Zhou, and L. Jiao, "Adaptive super-resolution for remote sensing images based on sparse representation with global joint dictionary model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2312–2327, 2018.

[7] Z. Gao, L. Ding, and C. Xiong, "Single image interpolation using texture-aware low-rank regularization," *Chinese Journal of Electronics*, vol. 27, no. 2, pp. 374–380, 2018.

[8] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[9] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[10] S. Izadpanahi and H. Demirel, "Motion based video super resolution using edge directed interpolation and complex wavelet transform," *Signal Processing*, vol. 93, no. 7, pp. 2076–2086, 2013.

[11] W. Chen, Q. C. Tian, J. Liu, and Q. Wang, "Nonlocal low-rank matrix completion for image interpolation using edge detection and neural network," *Signal, Image and Video Processing*, vol. 8, no. 4, pp. 657–663, 2014.

[12] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891–906, 1991.

[13] D. Sun, Q. Gao, and Y. Lu, "A MAP approach for image deblurring based on sparsity prior and Laplacian mixture modeling," in *2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 901–906, Hefei, China, May 2017.

[14] W. Dong, L. Zhang, R. Lukac, and G. Shi, "Sparse representation based image interpolation with nonlocal autoregressive modeling," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1382–1394, 2013.

[15] D. Sun, Q. W. Gao, and Y. X. Lu, "Image interpolation via collaging its non-local patches," *Digital Signal Processing*, vol. 49, no. 2, pp. 33–43, 2016.

[16] D. Sun, *Image Interpolation Based on Fractal Self-Similarity and Non-local Patches Collaging*, Anhui University, Hefei, 2016.

[17] D. Sun, F. Li, and Q. Gao, "A dynamical analysis on non-local autoregressive model and its application on image reconstruction," *Chaos, Solitons & Fractals*, vol. 130, article 109427, 2020.

[18] M. Elad, *Sparse and Redundant Representation: From Theory to Applications in Signal and Image Processing*, Springer, 2010.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8816760, 15 pages
https://doi.org/10.1155/2020/8816760

Research Article

System Design for Opportunistic Spectrum Access Using Statistical Decision-Making and Coded-MAC

Enrique Rodriguez-Colina, Ricardo Marcelín-Jiménez, Leonardo Palacios-Luengas, and Michael Pascoe-Chalke

Department of Electrical Engineering, Autonomous Metropolitan University, Iztapalapa, 09340 Mexico City, Mexico

Correspondence should be addressed to Leonardo Palacios-Luengas; lpluengas@gmail.com

Received 28 March 2020; Revised 29 July 2020; Accepted 7 October 2020; Published 23 October 2020

Academic Editor: Yin Zhang

Different mechanisms have been proposed to solve opportunistic spectrum access (OSA). In order to address spectrum management efficiently, these mechanisms can be divided into four main functionalities, spectrum sensing, decision-making, sharing, and mobility. These functionalities depend on the interpretation and adaptation of different parameters, for example, sensing and data interpretation for adaptive modulation, power adjustments, and changes regarding the range of frequency operation. For the decision-making function, a novel approach is proposed in which coding information is added to the establishment of the communication process thus assisting the medium access control (MAC). The presence of cognitive radio devices in the network coverage range can be controlled or coordinated by using specific redundancy codes. Hence, Reed Solomon (RS) code is used in this paper as part of the handshaking process to provide error correction. In addition, a redundancy strategy based on Rabin's information dispersal algorithm (IDA) is presented to provide fault tolerance to the communication between cognitive radio devices. In this case, the information is divided into fragments dynamically, and each fragment is coded by an RS code and reassigned to a subset of recipients using alternate paths. This work shows how to optimize spectrum access based on IDA and RS codes to diversify channel occupation without losing significant information with several frequency hops presented in cognitive radio communications. The validations were executed in a discrete event simulator developed in Python. The proposed system for OSA was found to perform better than other approaches using pilot sequences. Our proposal, therefore, provides fault tolerance, to diversify channel occupation, and helps identify the presence of primary and secondary users when a common control channel (CCC) is implemented by the optimization of the spectrum use.

1. Introduction

The implementation of a practical and useful cognitive radio (CR) device that attempts to achieve opportunistic spectrum access (OSA) requires several functionalities that using current technology results in various open issues to be solved. In the last 15 years, there has been exhaustive research to solve the main issues to make CR networks (CRN) effective and practical [1–5]. Several authors have proposed different functionalities for CR devices which can be divided into four main functionalities, such as spectrum sensing [6, 7], decision-making [8], sharing, and mobility in order to solve the OSA efficiently [9]. These functionalities depend on the interpretation and adaptation of different parameters, for example, sensing and data interpretation for adaptive modulation, power adjustments, and the changes regarding the range of frequency operation.

In the proposed system, the CR device is able to work with current available technology and avoids the use of a dedicated common control channel (CCC). This can be achieved mainly through changing its operation parameters in four domains: space, frequency, time, and coding. The decision-making functionality in combination with coding to medium access control (MAC) is used to identify cognitive devices in the network and can be implemented for the selection of unused bands. The presence of cognitive users is coordinated by the use of redundancy codes to cause minimum interference to the primary users (PUs) during the CR devices'

handshaking process. Furthermore, the CR device considers that sensing functionality is performed accurately and as fast as possible. The ability of each cognitive device to modify its parameters accurately and as fast as possible is directly related with interference avoidance to primary users. In addition to coding, information redundancy is considered. Here, the dispersion process considers fragments of n information that are sent through different channels. Thus, due to constant channel interruptions, it will be possible to recover the information with only m surviving dispersals. Observe that Figure 1 shows the domains to change communication parameters.

The four main functions that the CR device should perform are based on obtaining knowledge from the environment, making decisions using the information acquired, adapting parameters to modify operation, and learning from other devices. The sensing function obtains information from the environment and spectrum sensing is performed by the CR device. For this purpose, the spectrum can be divided into frequency slots which correspond to communication channels or frequency bands to be sensed. Spectrum segmentation in slots facilitates working with digital numbers or codewords in order to describe whether the channel is free to be used or occupied. A time-division approach is used to determine whether other CR and PUs are operating in the same frequency slot. Using this time-division approach, the presence of PUs is detected, and then, the CR devices change their communication channel without losing contact with their communication peer and with a minimum time interference to PUs. This is mainly performed through sending out predefined set of previously selected channels from the proposed decision-making module (DMM), which in turn requires the use of packets with specific information in the header frame.

The CR devices can operate in different environments where either primary users can exist or not. For the establishment of connection and channel mobility, the CR devices basically operate under the sharing technique for spectrum access known as underlay [10], in which wideband and low power signal below a predefined threshold is used for establishing connection, i.e., handshaking and mobility tasks in the presence of primary users. During the communication process, once communication has been established, an overlay technique is used in which interference to primary users is only reduced to the short time that it takes to detect the primary user signal and in addition to the time it takes the CR device to hop to another channel.

The remainder of this paper is structured as follows. Section 2 presents a state-of-art review, focusing on related works. Section 3 discusses the system's main contributions. Section 4 describes the proposed system which includes the significant functionality modules, such as slots for spectrum sensing, decision-making, and the Coded-MAC Algorithm. Section 5 presents a simulation and analysis of the results. Finally, section 6 provides the conclusions.

2. State-of-Art Review

Cognitive radio (CR) technology is a paradigm that promises to be used in the future for wireless communication systems. For this purpose, different strategies have been developed which can be divided into four functionalities: (i) detection, (ii) decision-making, (iii) exchange, and (iv) spectrum management mobility. This section considers some strategies for DMM used in CR devices and will basically focus on network coding [11, 12].

Some strategies are based on coding for information exchange between a base station (BS) and users. Zhao et al. [13] propose an asymmetric network coding. The main idea is to apply the concept of CR in network coding transmissions, in which BS tries sending new information while helping user transmissions as a relay. Qu et al. [14] proposed a study on learning how to use two-dimensional multiarmed bandit (MAB) to use network coding to improve SU throughput in CRNs when channel quality is unavailable at SUs. Qu et al. [15] developed an efficient network coding strategy for SUs while considering uncertain idle durations in CRNs. Essentially, systematic network coding (SNC) is employed to opportunistically use the idle duration left by PUs. Liang et al. [16] proposed an adaptive dynamic network coding scheme (ADNC) conceived for cooperative cognitive radio (CCR) for devising a novel ADNC-CCR system. Qin et al. [17] investigated the limitations of transmission control protocol (TCP) in multichannel multiradio multihop CRNs and proposed a novel TCP called TCPJGNC (TCP Joint Generation Network Coding, JGNC) based on network coding. Khosroazad et al. [18] proposed a physical layer network coding an extended method for CR, the main goal of which is to maximize CRN capacity while keeping the total interference imposed on the primary users under a certain threshold.

Other techniques are based on the improvement in spectrum exchange over routing channels. For example, Zhu et al. [19] mentioned that OSA is one of the core technologies of cognitive radio, which emphasizes the intelligence of the network and adapts spectrum utilization, and in this publication, the authors consider the analysis of each channel to evaluate the best option. We do the same but the difference consists in the method used which provides a proactive approach to the decision-making process combined with a reactive approach. The main improvement is that we provide a solution algorithm to distribute the communication in the spectrum opportunities, (i.e., spectrum holes) by using backup channels dynamically assigned and which contain the dispersed information with the use of Rabin's IDA algorithms. A crucial difference in the decision-making process in our proposed system is to consider contention rather than a time slot operation or cooperative approaches as most of the literature shows, this a more practical point of view since most of the systems operated with contention for medium access.

Sangi et al. [20] proposed a channel-route failure-based Cognitive-Ad-hoc on-demand distance vector (Cognitive-AODV) routing protocol with the modifications in channel-route-error (channel-RERR) to detect the exact channel-route failure and provide the best alternate end-to-end channel-route path in between source and destination. Deng et al. [21] proposed a cognitive routing and optimization protocol based on multiple channels with a cross-layer design to study joint optimal cognitive routing with maximizing

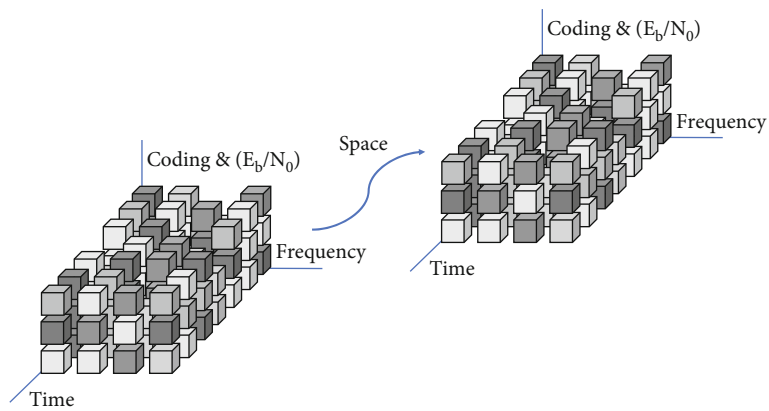


Figure 1: Four domains for adaptive parameters. The blocks represent current communications for the time, frequency, space, and coding parameters.

network throughput and network lifetime. Mei et al. [22] proposed a hybrid network coding scheme under different channels to improve the efficiency of idle spectrum utilization, the successful transmission rate, and the packet loss rate that can be increased and decreased through redundant coding. Sboui et al. [23] proposed a new approach using a multilayer-coding (MLC) strategy, i.e., a broadcast approach, to enhance spectrum sharing over fading channels.

Other approaches are based on resilience to the interference caused by PUs and noise disturbances via information dispersal in SU's messages over multiple frequencies and time intervals [24]. Hamza et al. [25] proposed a new multiuser transmission coding scheme, cooperative quadrature physical-layer network coding (CQPNC) for cognitive wireless networks.

Nadendla et al. [26] proposed a mechanism using the information dispersal algorithm (IDA) and Reed Solomon (RS) code. Hence, with Rabin's dispersal algorithm, the confidentiality and integrity of the messages are evaluated. For this purpose of redesigning the original version and evaluating the proposal, the presence of an attacker is considered as an attempt to compromise the communication paths because there is an intentional breach in the confidentiality and integrity of the source message. In addition, the RS code is used as an error corrector. In this paper, a DMM based on coding and dispersal information algorithm is proposed. The main contributions are described in the next section.

3. Main Contributions of the Proposed System

This work's main contribution is the development of a DMM in which coding is added to the establishment of the communication process. The DMM that was developed makes use of an RS code that offers low processing power requirements and a considerable amount of advantages to identify encoded signals intended for wireless applications. Thus, one of the main advantages of our mechanism is that it communicates with an RS code to establish the communication that assigns an identification code used for an additional connection in the communication range [27]. This makes communication robust to noise and able to compensate losses caused by multiple frequency hops. In this way, adequate effectiveness in the coding rate must be achieved; for this purpose, different RS coding simulations were carried out considering that very short bursts of information must be transmitted and have a high processing speed. Thus, for our proposal, an RS codeword (15,9) was implemented, which considers an adequate coding rate according to different functional tests.

An additional consideration is that PU operation is not predictable in most cases and CR users cannot obtain a reliable communication band. Moreover, CR devices may not detect any single spectrum band to meet application requirements. CR users can adopt a multiradio transmission method in which each radio interface transmits to different spectrum bands. Transmission in multiple spectrum bands, i.e., diversity, allows lower power to be used in each spectrum band. As a result, primary users could support multiple spectrum selection capabilities for transmission, thus requiring the use of distributed system techniques to manage redundancy. If communication experiences losses due to mobility, for instance, the information transmission will fail. The most frequent solution to this problem in a noncognitive network consists of sending a backup copy of the information through a second device, or an alternate path. This solution's main drawback is that the alternate component reduces its effective available performance in order to manage the backup copy. Besides, if the first and second components fail, the transmitted message will be lost. Hence, a common feature is the use of a redundancy strategy in order to provide fault-tolerance and, particularly, integrity. In this context, the simplest approach consists of file replication. Due to CR characteristics, it is necessary to consider a proper combination of the number of disperses to be generated per file. For this, the probability of failure is considered. As a result of this analysis, with the increase of disperses, the probability of failure occurrence increases, thus implying longer dispersion time and recovery. The DMM has the following characteristics:

(i) The proposed RS code mechanism makes communication robust to noise and able to compensate for additional losses caused by multiple frequency hops

(ii) Using RS codeword (15,9), an adequate coding rate is achieved with low power for the process of exchanging handshaking between cognitive devices

(iii) A redundancy strategy based on Rabin's information dispersal algorithm is presented to provide fault tolerance. Here, the information is divided into fragments, each of which is coded by an RS code and reassigned to a subset of recipients or an alternate path

(iv) A configuration of IDA with parameters (5,3) is selected, considering an adequate relationship according to different analyses. Thus, the proposed system sends five dispersed fragments through different channels and, in case information is interrupted, the information can be retrieved with only three of them

(v) A solution algorithm to distribute the communication according to the spectrum holes availability by using backup channels dynamically assigned with information dispersed with the use of Rabin's IDA algorithms

(vi) We proposed a method for the DMM which provides a proactive approach to the decision making process combined with a reactive approach which improves the response of the system

(vii) The proposed system considers contention for the MAC rather than a time slot operation or cooperative approaches, this a more practical point of view since most of the systems operated with contention for medium access

4. System Description

A CR using currently available technology was designed, focusing on four main characteristics in order to as much as possible avoid the interference to PUs combining underlay and overlay approaches and thus have a robust decision-making mechanism based on statistical analysis, i.e., proactive and reactive decisions [28, 29]. This aims to facilitate the coordination of CR devices and reduce the computational complexity for required control instructions and coding. The proposed model comprises five main modules to perform the OSA functionalities and effective communication, as shown in Figure 2. The system's input requires a collection of sensors which represent the Physical Interface. As aforementioned, spectrum sensing is very demanding and requires the highest possible sensitivity.

4.1. Spectrum Sensing (Monitoring). Most of the testbeds developed operate within the 2.4 GHz band. Thus, an example of a free licensed band is provided in which the frequency spectrum is divided into frequency slots which correspond to the spectrum frequency channels. Consequently, dividing the spectrum into slots facilitates working with discrete numbers and enables a representation of the occupied and free channels as a vector of logic ones and zeros. Figure 3 illustrates how the 2.4 GHz band is divided, for example, into 14 channels which can be seen as spectrum frequencies with the presence of a primary signal and secondary signals. Channels size is independent of current mobile technology, i.e., it can be used in any part of the frequency spectrum.

Segmentation of the frequency spectrum serves to increase resolution for sensing capabilities (see Figure 4). This segmentation allows for the determination of the presence of energy in a reduced frequency band in order to detect the presence of PUs more accurately. With the segmentation process, more sensors and computational power are required in order to avoid interference. However, for our analysis, the result segmentation with occupied and free channels is taken as a vector of logic ones and zeros.

4.1.1. Decision-Making Module (DMM). Several factors and attributes should be considered to make a proper decision based on sensed or captured information. The proposed DMM comprises two main capabilities: the first is based on a modified version of the analytic hierarchy process (AHP) algorithm, and the second uses assigned codes for a MAC mechanism which allows an easy identification of CRs. The second module provides significant information about the status of the CR device, such as the power range and the detection of nearby PUs. As a powerful tool to avoid interference during the handshaking process, the use of a Reed Solomon (RS) Code in the MAC mechanism was proposed. This code can be implemented without the requirement of complex computing capabilities on the CR device. This code is also used as a CR device identifier in order to increase fairness among secondary users accessing the communication medium. The information received by the CR device is used for local and distributed cooperative decision-making.

4.1.2. Multiple Criteria Decision-Making Module. The AHP is part of the proposed multiple criteria decision-making module. The AHP is a general problem-solving method that serves to make complex decisions, for example, multicriteria decisions based on variables without numerical costs. The decision makes modeling through using multiple attribute AHP which consists of the following five steps: structuring the decision model, entering alternatives, establishing priorities among elements, synthesizing, and conducting sensitivity analysis. In the case of our proposed decision method, the AHP is complemented and combined by statistical analysis of the behavior of the channels stored in a database and correlated with the last measurements "snapshot-measurements" and then to provide feedback data to the AHP as an additional element for the decision. This considers the statistics through time of the channels, i.e., the history of the channel behavior is provided as feedback to the system to be combined with AHP, as shown in Figure 5. The AHP organizes the decision-making breaking down the decision problem into a hierarchy of interrelated decision elements. At the top of the hierarchy lies the main decision objective, which in this case is to select the best band available. The lower levels of the hierarchy contain attributes, i.e., objectives which contribute to find the best solution for specific characteristics or applications. At lower hierarchy levels, the detail increases, and once the hierarchy

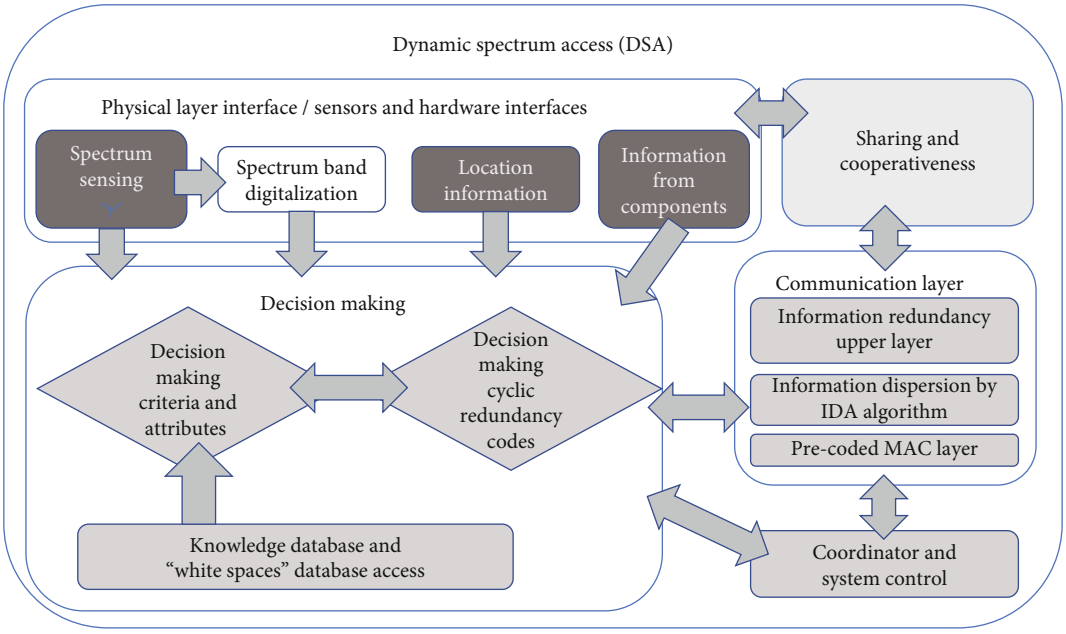


Figure 2: Simplified block diagram of the proposed CR device.

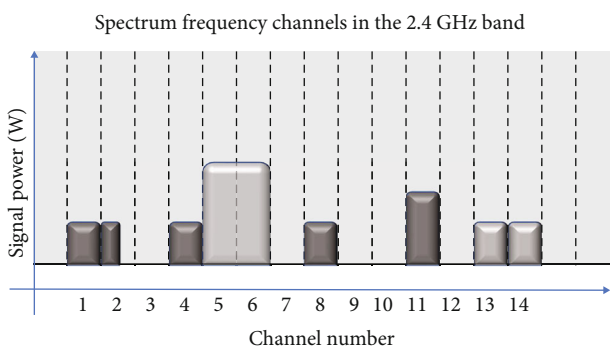


Figure 3: Frequency slots for 14 channels system with occupied and free channels.

is established, the CR devices evaluate the alternatives using the given criteria in order to obtain the priority vectors, which are also known as criteria weights. The attribute samples are periodically supplied to the decision-making function. This function processes them and produces a ranking of the frequency bands ordered from best to worst as a function of the criteria weights. Since there are numerous combinations for the spectrum bands between communication source and destination, it is infeasible to consider all possible bands for spectrum decision. However, the AHP model can manage a considerable number of bands without a significant delay, which is around 6 microseconds for 15 bands. Similar to [30], a modification was proposed which includes a feature with output-related feedback regarding the statistics managed by the modified AHP. Information feedback regarding the decision-making mechanism includes error probability. This feedback is a function of the four criteria previously provided to the system; consequently, a fifth criterion is added. This alters the rules of the original AHP, and therefore, the behavior of the decision-making mechanism is also altered.

Figure 5 shows the AHP feedback model referred to in this paper as the feedback AHP (FAHP) algorithm. In the model, each time the CR device performs the spectrum sensing function, it obtains snapshot-measurements from the environment. The snapshot-measurements are sent as attribute samples that feed the decision-making function. Here, FAHP is used to calculate the bands ranking taking attribute values into account. This band ranking is represented as a vector with the available options from the best to the worst bands for each sample, as depicted in Figure 5. Simulation results show that the proposed FAHP algorithm computes a trade-off between diverse priority attributes meeting the application requirements. The resulting band ranking, i.e., "processed sample" outcome is sent to the CR mobility and sharing functionalities and is simultaneously kept in the knowledge database for further processing. This database processing results in histograms characterizing the number of times that the bands have been selected for each position. Subsequently, the statistics for each frequency band are contrasted with the "processed sample vector" (SV) for each interaction of the decision-making process. This contrasting of information provides possible discrepancies between the "processed sample vector" and statistical behavior, which is in fact current ranking versus statistical ranking.

Hence, a probability error when selecting available options can be determined by comparing the statistical ranking with the current ranking vector. The "processed sample vector" is a parameter that indicates the transients in band ranking. In addition, the statistical database helps to identify possible periodical behavior for the band ranking. Although the AHP helps to find optimum bands for a particular application or service, a combination of decision methods should be essential for the CR communications.

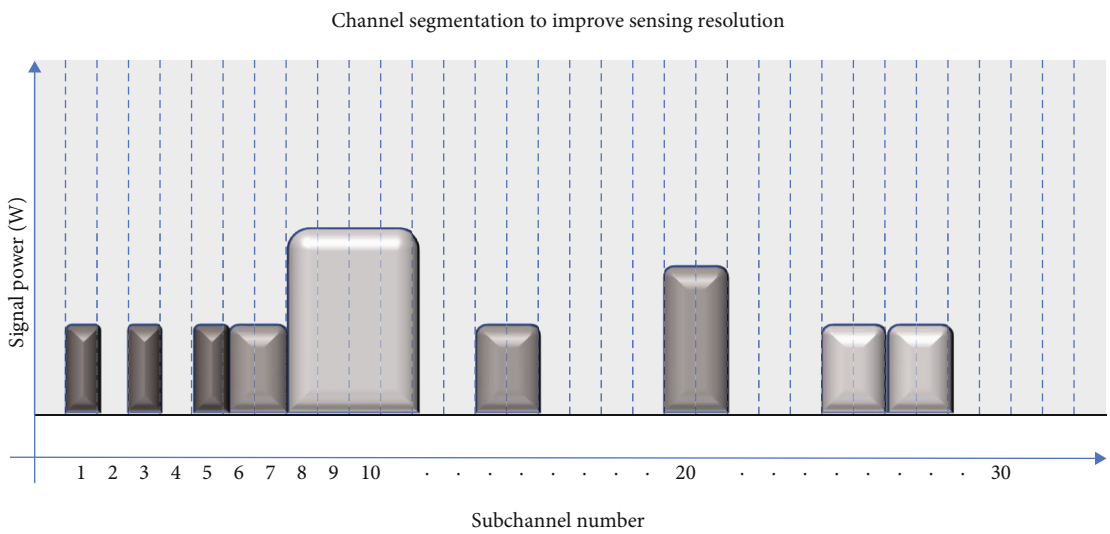


Figure 4: Segmentation of frequency slots to form subchannels.

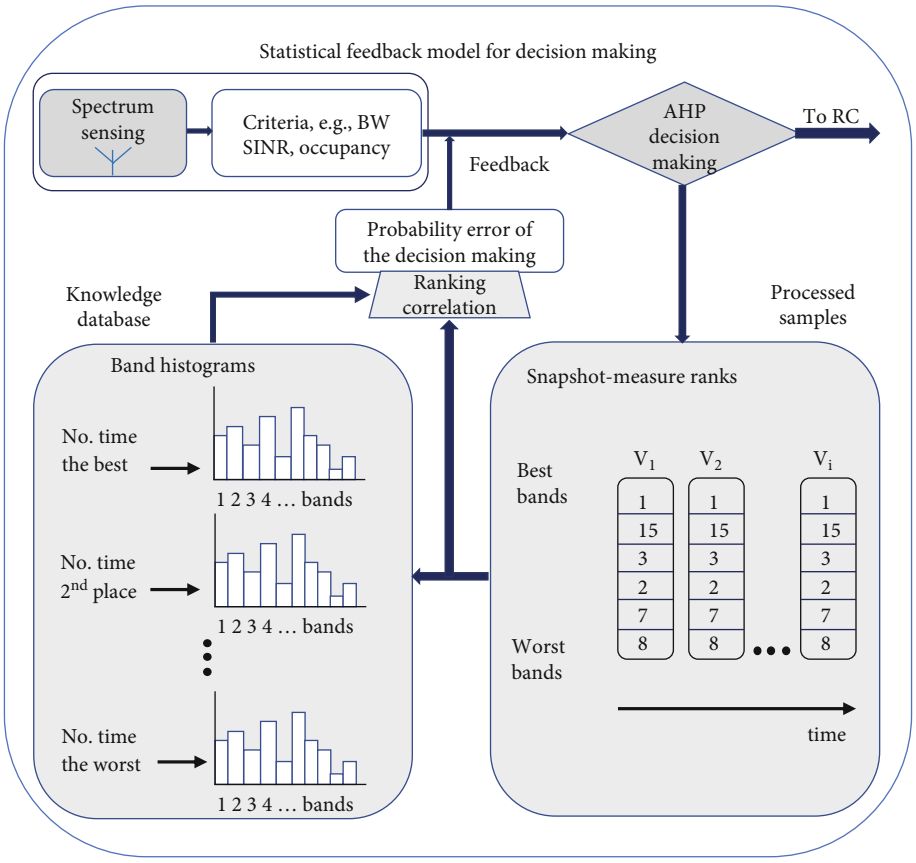


Figure 5: Proposed feedback AHP model.

4.2. Coded-MAC Design. The proposed Coded-MAC incorporates the RS code for the handshaking and mobility processes and IDA as an extra redundancy mechanism communication. The proposed model also comprises the use of an identifier for the communication between CR devices which improves the connection with a device without the need of a dedicated CCC. The Coded-MAC has been tested with a developed discrete event simulator to characterize CR devices communicating in the presence of PUs. Therefore, if the data is transmitted in burst mode or the data is sent through different routes, the RS code can provide error detection and correct them. This enables a robust

communication without interferences with the PU. Furthermore, it is suitable for the cognitive radio environment in which the packet loss is severely affected by the constant frequency hops. The RS code also works for synchronization between cognitive devices when movement to another frequency band is required. In addition, low power can be managed for the handshaking process among cognitive devices.

4.2.1. Reed Solomon Coding. Regardless of the coding used in the transmission systems, e.g., CDMA and OFDMA, adaptive forward error correction (FEC) is commonly used in wireless communications, such as convolutional, Reed Solomon, and Turbo codes [31–33]. In this paper, a novel application for the decision-making function is proposed in which a Reed Solomon (RS) code can be implemented for communication. Communication between secondary users can be controlled or just coordinated through using specific redundancy codes. Due to communication's constant intermittency, a short length of the data block and transmission speed should be considered. In order to evaluate block length and redundant information that should be added, a communication system was proposed using BPSK modulation and demodulation. Initially, the data is binarized and encoded. Subsequently, the encoded binary data is passed through a Gaussian channel in which Additive White Gaussian Noise (AWGN) is added to the channel. Afterward, the binary noisy data are demodulated and decoded. Figure 6 compares the probability of error obtained from uncoded messages and messages encoded with (z, k) RS code for different blocks such as (7, 5) RS, (15, 9) RS, and (20, 15) RS code. Note that error probabilities are presented as a function of the ratio of energy dedicated to a bit E_b and the spectral density of noise N_0. Thus, Figure 6 demonstrates that as the energy of the bits increases, the propagation of error decreases since the message is less affected by noise.

The different bit error rates (BER) are shown in Figure 6, the message encoded with the (20, 15) RS code, and the performance of the error rate improves with respect to the uncoded message. Regarding the (7, 5) RS code, it improves E_b/N_0 in order to reach a bit error rate of 10^{-3} in approximately 6 dB. However, (15, 9) RS code remains below in all cases, which improves the E_b/N_0 to reach a minimum BER of 10^{-8} in approximately 7 dB. Thus, the RS code used for the proposal is (z, k) in which the block length: $z = 15$ symbols, and the data length is k symbols = 9. The encoding algorithm expands a block of 9 symbols to 15 symbols by adding 6 redundant check symbols. The entire z-tuple space contains $2^n = 215 = 32768z$-tuples, of which $2^k = 29 = 512$ (or 1/64 of the z-tuples) are codewords. Consequently, the RS codes have the remarkable property that they can correct any set of $z - k$ symbol erasures within the block. A wireless device can easily compute these codes without significant power consumption.

According to Figure 7, the approach shown here is mainly planned as an infrastructure network topology in which a base station (BS) helps to connect devices in order to share information all over the network. This approach, however, is not constrained to infrastructure architectures, so it can be implemented in other types of architecture, such as Ad-Hoc networks. When a CR user enters a specific location area, it should identify itself as a CR device. Thus, the BS assigns a local identification to the CR device which is, in fact, a communication code apart from the device's identification (ID), independent from of the permanent address and physical ID number. The communication code assigned by the BS identifies the device as cognitive radio and in our proposal is transmitted using the RS code. This RS cognitive code (RSCC) identifies the device at the location and power range of the BS and identifies only cognitive device communications. The RSCC helps to gain fair access to the CR device network because this registration process distinguishes the CR device from other opportunistic users in the area, and

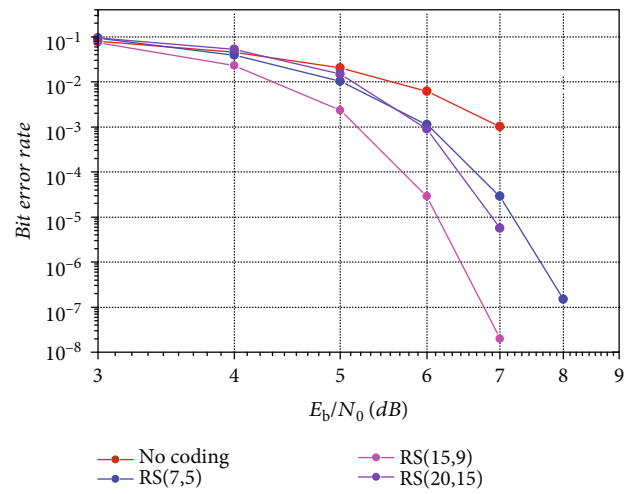


Figure 6: Comparison of bit error rate performance of (z, k) RS code for different block lengths.

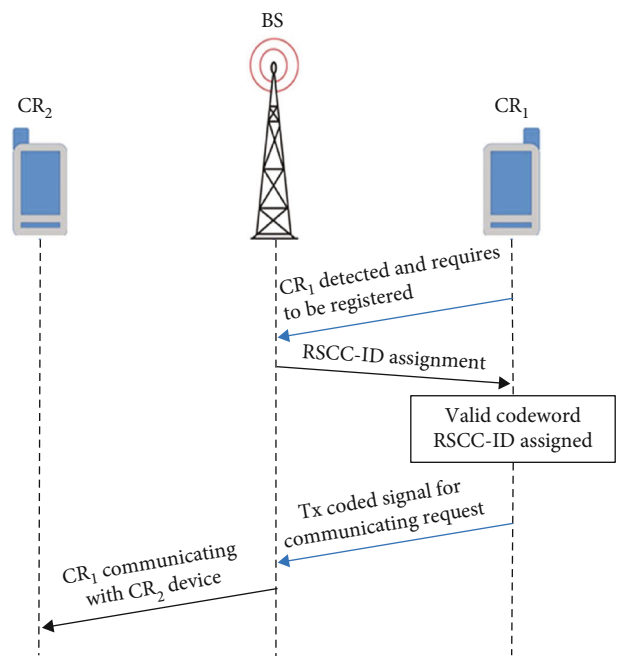


Figure 7: Handshaking process for two CR devices.

consequently, a more effective media access control (MAC) can be implemented.

Once the CR device is registered on the network and when this device is communicating, the RSCC identification code also serves for the reception of poor signal detection. In addition, with the use of RSCC during communication, privacy and recognition of the communication signal are managed by the CR device and the BS. The CR device signal can be identified and differentiated from primary users and other devices. Another variant of this process provides codes for primary users making it easier to identify between primary, secondary, and other types of communication devices in the area. Thus, the proposal is to use the (15, 9) RS code in order to determine the RSCC for identification of the CR devices, which was implemented in the simulation using a developed program in Python.

4.2.2. The Proposed Coded-MAC. The proposed Coded-MAC solves the opportunistic spectrum access problem considering an overlay approach when there is a pair of CR devices communicating with each other. However, it uses an underlay approach when the handshaking for the connection of two CR devices is in process. This eliminates the use of a CCC, although minimum interference is produced during connection.

The interface lapse depends on the spectrum sensing time and the exchange of messages to coordinate common free channel options. The spectrum sensing's function is to inform the CR devices about spectrum white spaces. Once the connection is established on an available channel, the interference is null until a PU attempts to occupy this free channel. The list of available channels is updated according to changes in the spectrum; hence, if a primary user enters that portion of the spectrum, a spectrum mobility mechanism is set off based on the previous sensing information of the backup (BK) channels. The sensing information of channel availability is acquired continuously and stored. To deal with the "hidden terminal" and "the far-away from terminal" effects, the proposed Coded-MAC is designed in response to the frame exchange sequence (FES), RTS-CTS-DATA-ACK. The sequence of received acknowledgments is also controlled. The CR device that requires to communicate mounts its signal as a peer to peer network connection on a free channel, using a network identifier (idNetwork).

This idNetwork, in addition to the CR identifier, is used by the caller CR device in order to specify to whom it requires to call. Then, the called CR device senses the spectrum and knows that another CR is calling. The Coded-MAC also comes up with a solution when CR devices cannot detect the same available channels by sending coordination packets to establish communication in other free channels by emitting 4-bytes coordination message. This coordination message produces a reduced interference time to the PU that has been coded, which is equal to the four bytes divided by the data rate. The proposal considers network scalability, and consequently, a modified version of the overlay carrier sense multiple access with collision avoidance (O-CSMA/CA) protocol for CR is implemented.

Figure 8 shows the frame used for general communication between CR devices. The first block of the frame includes the information necessary to start the handshaking process and to identify whether a packet has been sent successfully or not. Then, the idSource is the CR identification of the caller device, and the idTarget is the identification of the CR device to be called. The channel transmission block determines the channel through which the pair of devices must communicate. The next block indicates the BK channels used to move the communication if the channel transmission becomes busy. These channels are used in case of the presence of a PU and when the CR devices must change their transmission to another free channel. This action is referred to as "channel hops." The counter block is used to count the number of attempts through which to find a free channel to communicate. The idNetwork is a parameter which identifies the communication in progress. The idNetwork is also used when the CR device wishes to call another CR device. Then, a communication signal with the idTarget and the idNetwork is sent through an available channel until the required CR device receives this signal and answers the call. This establishes the communication in the available channel. During the handshaking (see Figure 9), there are two possible options: the CR devices move to a free BK channel, if available, or they search for a new channel through which communicate and are obliged to wait. In fact, the idNetwork is an identifier number which corresponds to the signal of the CR device trying to communicate, so the called device is able to identify by scanning the spectrum who is calling and to acquire the list of available channels for both devices. Thanks to the idNetwork, the CR transmitting can identify whether the accessing device is a PU or another CR. The block in the frame with the amount of packets indicates how many packets have been successfully sent to the receiver. Finally, the data stream is incorporated in the last part of the frame, in which the information is divided into fixed-size packets.

4.2.3. Communication Stage. Figure 10, which shows the flow chart for the handshaking and communication between CR devices, illustrates the OSA Communication process stage. The following considerations should be assumed. (i) Each CR device has an associated identification number (id). (ii) It is assumed that communication involves CR devices, sources (CS), and targets (CT). (iii) Redundant channels must be considered for use in case a PU arrives during a communication. The communication process starts the moment the CS sends a call request. The following selection criterion is applied: when communication is established the CR network with a lower id is chosen. This connection is made in the same channel, and it is necessary to choose only one. If this condition is met, CT establishes a connection in the free channel using an id network. Consequently, reading the identification network created for the CS, the devices continue to detect the medium during communication to verify if the channel is still free or whether the PU has arrived. If at any time a collision is detected (the entry of a PU has been detected), CR devices move as fast as required given the BK channel information they have. According to Figure 10, if the CR devices cannot find a common free channel, the CS

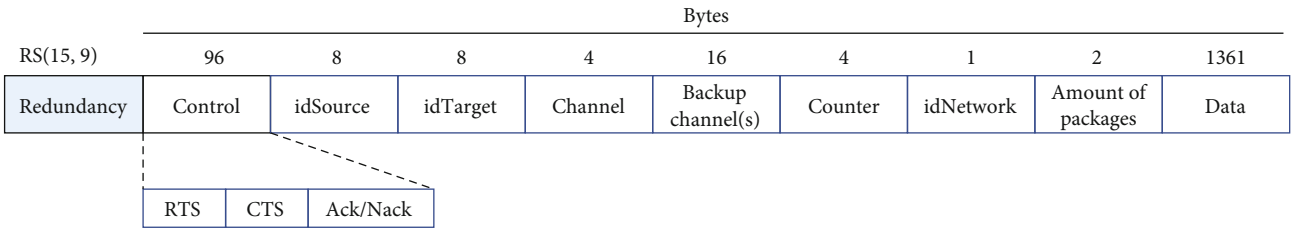


Figure 8: CR communication frame in Coded-MAC.

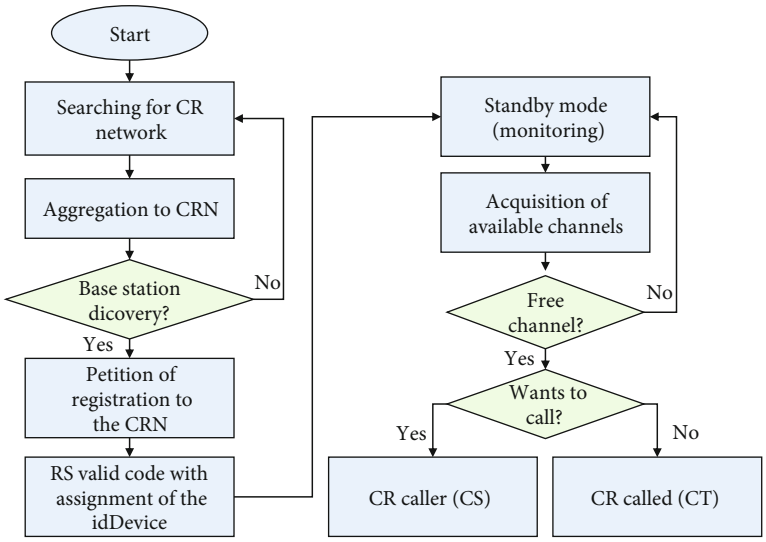


Figure 9: Network registration and communication.

cancels the call and repeats the process. On the other hand, CS checks whether the TC is communicating through the new free channel in order to continue with the communication interrupted by the PU. For this, the system has a counting mechanism that indicates the number of attempts made. If the number of attempts is smaller, then the CS updates the network's id and waits for the CT's response. In another case, the handshaking process starts from the beginning.

4.2.4. Additional Redundancy Stage. If communication experiences losses due to mobility, the information transmission will fail. The most frequent solution to this problem in a non-cognitive network consists of sending a backup copy of the information through a second device (or an alternate path). This solution's main drawback is that the alternate component reduces its effective available performance in order to manage the backup copy. Besides this, if the first and the second components fail, the transmitted message will be lost. Hence, a common feature is the utilization of a redundancy strategy in order to provide fault-tolerance and, particularly, integrity. In order to send a file, it can be split into various information units. In this case, several copies of the resulting units are allocated to a given subset of devices. In contrast, there are systems in which information redundancy is implemented through using error-correcting codes [34]. The use of an additional approach to the Reed Solomon code based on the Information Dispersal Algorithm (IDA), developed by Rabin [35], is here proposed. Let us say F is a file, F could be transformed into n files called dispersals. Each of size $|F|/m$, where $n > m > 1$. The dispersals are then sent to different channels. From the algorithm properties, it is granted that if any $n - m$ dispersals were lost, the original information could be reconstructed from the m surviving dispersals. In this case, the cost of the reconstruction is compensated due to the greater tolerance that the system supports, in addition to the reduction of devices required for redundancy.

Due to the features characterizing of cognitive radio, it is pertinent to think about what could be considered a good combination of m and n. For a fixed $K = n - m$, it is possible to have different combinations of n and m, but although $m \longrightarrow n$ would seem a good option especially when n is greater than or equal to 10, because redundancy is significantly reduced, this implies two considerations: (i) dispersion and recovery times increase, and (ii) what is even more important, the probability that K or more faults occur increases with n, as argued below. Consider two instances of IDA with parameters (5, 3) and (10, 8), respectively. Both offer the same tolerance, but the second case does so with a lower cost in redundancy (0.666 vs. 0.25, respectively). It would seem that the second case is

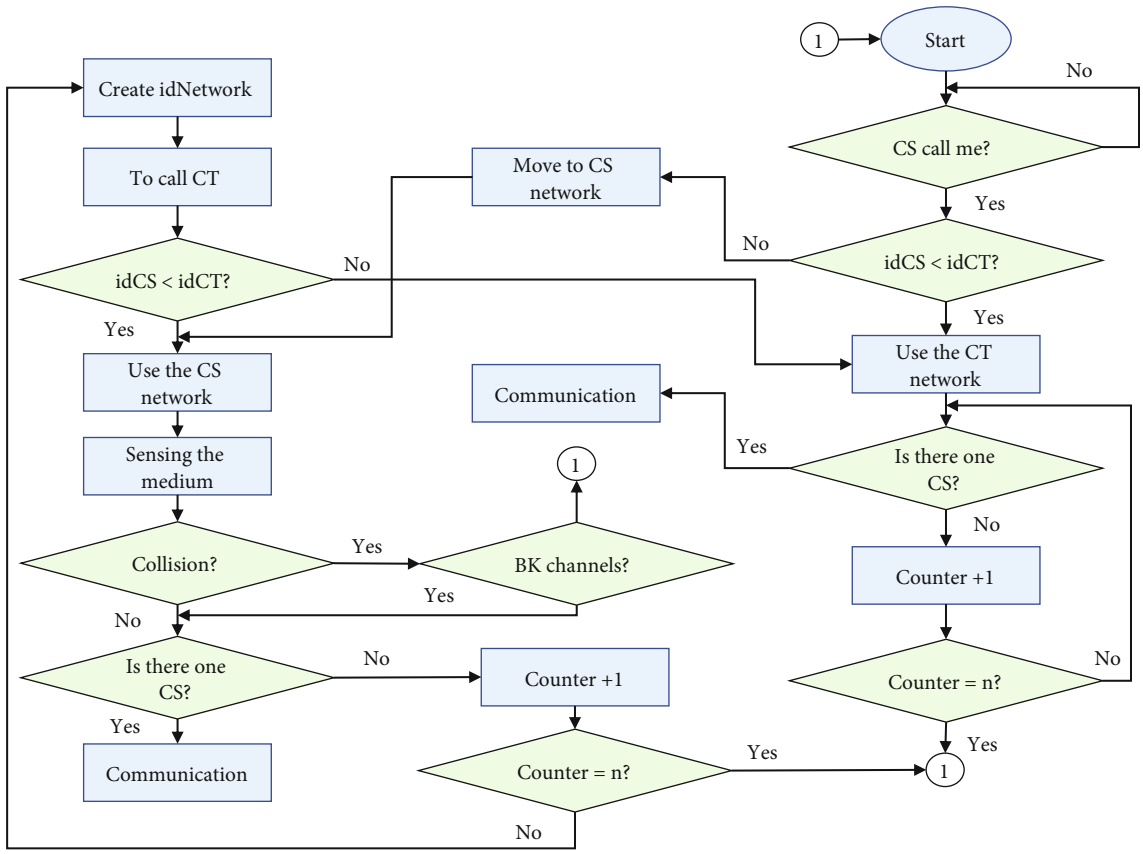


Figure 10: Communication, caller, and called CR devices.

better, but the probability that there exist two or more failures as a function of n should be considered. This is obtained by modeling the probability of faults with a binomial distribution where p is the probability that dispersal is lost since there are n initial dispersals, then the probability that two or more faults will occur should be calculated, but this is the complement that 0 or 1 failure occurs. The result is a function that grows with an n value. The conclusion is that, although it seems interesting to increase n, maintaining the number of failures that can be tolerated $(n - m)$, this criterion can lead to greater risk. It is also important to consider that a large n value implies a longer dispersion and recovery time. Therefore, alternative $m \longrightarrow n$, when n takes a very large value, e.g., >10 it does not seem very convenient given the processing cost but, above all, the risk faced by the overall system.

In this paper, (5, 3) was chosen. This means that an initial file is transformed into five dispersals and can be recovered with any three of them. Each dispersal is the size of $|F|/3$, and therefore, there is an excess of information, or information redundancy, equal to $2|F|/3$. Table 1 shows all possible combinations of IDA parameters for $n = 2, \cdots, 7$ and $m = 1, \cdots, 6$. Each entry represents a 2-tuple, (n, m), with its corresponding redundancy $((n - m)/m)$. Reading this information by rows, the rightmost entry represents the combination supporting the biggest number of losses or missing dispersals, for a fixed n. Nevertheless, the price to pay is an excess of redundant information, i.e., BW which, in the extreme case $(n, 1)$, implies that each dispersal is actually a replica of F. Reading by columns instead, each column represents all the possibilities that may compensate the same number of losses, i.e., the ith column (from left to right) describes all combinations that support up to i losses. It should be noted that a given entry supports the same number of losses as any lower entry in the same column. However, this results in high redundancy [36]. Figure 11 shows that each dispersal is sent through different channels, and in case of information interruption, the information with three dispersals can be recovered.

Table 1: IDA parameter combinations and their redundancy levels.

(2,1)1					
(3,2)1/2	(3,1)2				
(4,3)1/3	(4,2)2/2	(4,1)3			
(5,4)1/4	(5,3)2/3	(5,2)3/2	(5,1)4		
(6,5)1/5	(6,4)2/4	(6,3)3/3	(6,2)4/2	(6,1)5	
(7,6)1/6	(7,5)2/5	(7,4)3/4	(7,3)4/3	(7,2)5/2	(7,1)6

5. Tests and Analysis of Results

The results show a test scenario where PUs accesses the channels randomly. For this test scenario, if a pair of CR devices

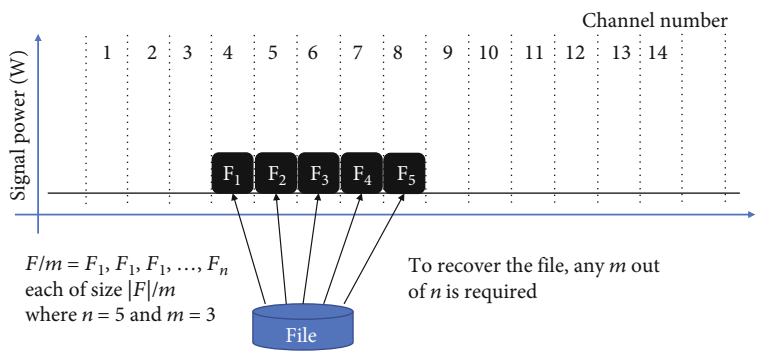


Figure 11: Use of IDA to send information through different channels.

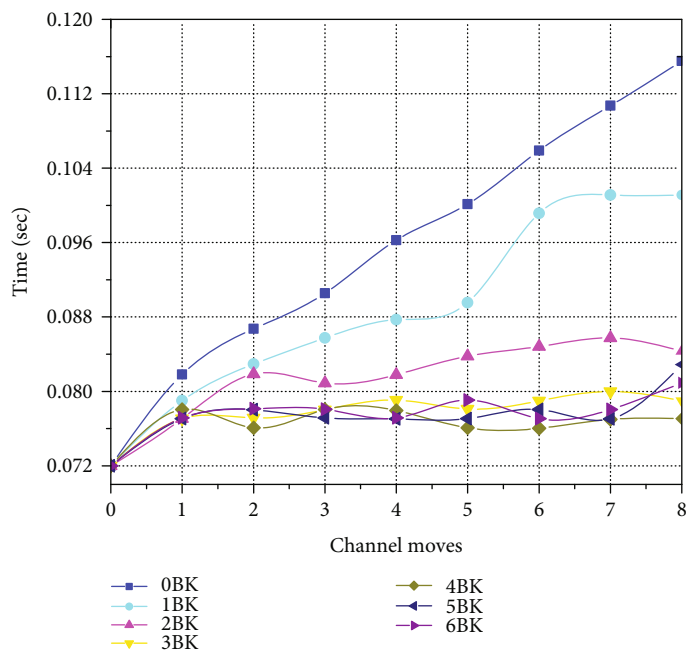


Figure 12: System performance with a pair of CR devices communicating, and with BK channels.

are communicating and a PU unexpectedly occupies the channel, the involved CR devices have to move as soon as possible to another available free channel. The simulation parameters are 5 Mbps data rate for the transmission of a file of 30 1500-byte packets. The file was split into packets in order to acknowledge the received packets and manage the number of correctly received packets, i.e., goodput. Thus, if the CR communication is interrupted by a PU, the CR communication can continue from the last acknowledged packet, although the CR devices are obliged to move to another channel. As aforementioned, a pair of CR devices establish communication while they are interrupted by a PU occupying the channel at random. This random occurrence simulates when CR devices move spatially from one place to another or just because the PUs operate switching on and off randomly.

5.1. A System with Two CR Devices Communicating. The CR devices exchange information by sending a testing file of 45 kbytes which is divided into 1500-byte packets. The optimum time to transmit this testing file without interruptions, i.e., without channel movements, is around 0.07 seconds, as shown in Figure 12. This is an optimal time because there is no need to move to another channel during transmission of the entire testing file. The time elapsed comprises the time of 30 packets sent and their corresponding acknowledgment received by the sender. Figure 12 illustrates the elapsed time during which the two CR devices are communicating and the number of channel movements is varied up to eight times. It also shows the communication when the number of BK channels varies from one to six.

The system's performance is depicted when BK channel availability is not always present. These results demonstrate when BK channels can be occupied by a PU in any unpredictable moment. The Coded-MAC performance was also tested for different packet sizes in order to find the most favorable ratio between channel hops and retransmissions of unacknowledged packets (see Figure 13).

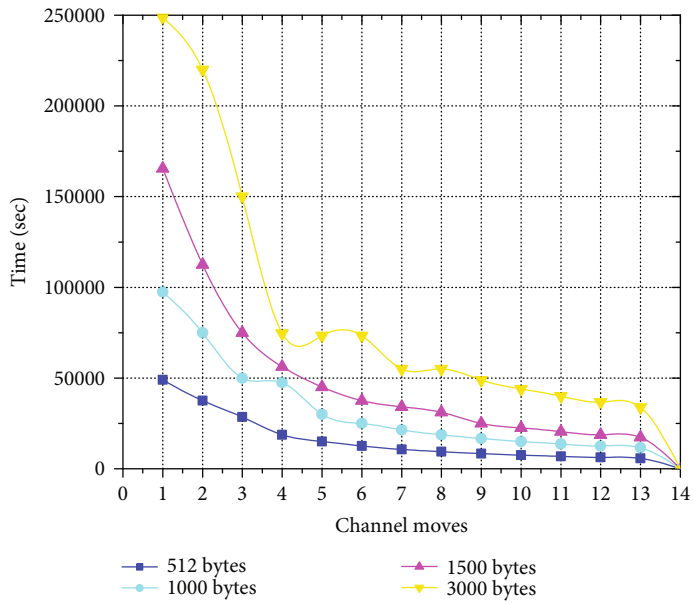


Figure 13: System performance varying the packet size.

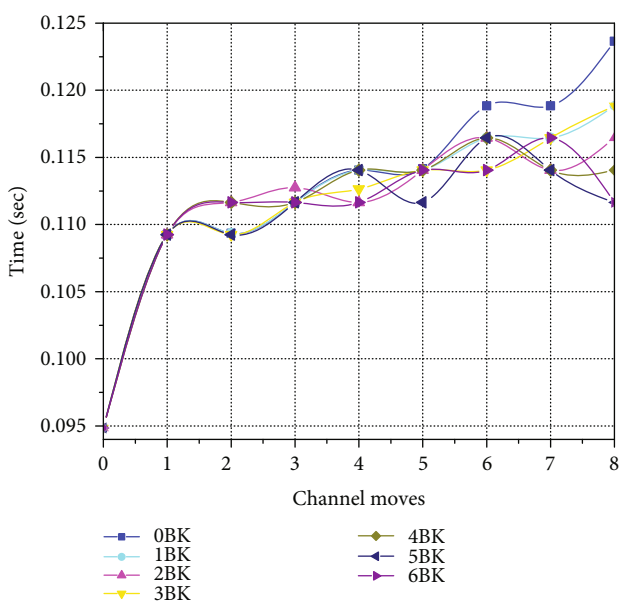


Figure 14: System performance with 20 CR devices communicating and with BK channels.

The goodput for the transfer of the testing file which is divided into 3000-byte packet sizes results in the best performance, as demonstrated in Figure 13. This is because the number of acknowledgments is lower than the other case scenarios in which packet sizes are smaller. This behavior was expected, but it was important to measure how the transmission would be affected by the interruptions per channel hops.

5.2. System with More than Two CR Devices Communicating. In the second scenario, 20 CR devices were considered. They each sent a 45-kbyte file and split it into 1500-byte packets at a 5 Mbps data rate. The communication was also evaluated when the number of BK channels varied from one to six. The random entrance of PUs into the channels occupied by CR devices was simulated with a random uniform distribution. The number of channel hops required by the CR devices was plotted against the time elapsed to conclude the testing file transfer. Figure 14 demonstrates that the elapsed time increases when the CR devices have to find free channels in order to establish communication. The proposed Coded-MAC performance was tested for different packet sizes in order to find the most favorable ratio between channel hops and retransmissions of unacknowledged packets. The simulation considers n fragments of the information sent through different previously chosen channels were similarly considered, and messages were able to be retrieved with only m dispersed. This causes information to be transferred quickly in small blocks, thus reducing energy consumption in CR devices.

5.3. Coded-MAC vs. O-CSMA/CA. Coded-MAC vs. O-CSMA/CA adapted for cognitive radio networks for the transmission of a file in which secondary users are contending for medium access were compared.

Figure 15 demonstrates that Coded-MAC performs similarly to the O-CSMA/CA when CR devices contend for the medium without PUs. However, if BK channels are available, the Coded-MAC has a significant advantage over the O-CSMA/CA.

As observed in Figure 16, the decision-making process is faster when backup channels are used to distribute information by means of dynamic dispersal of the information using Rabin's IDA approach and the Coded MAC that provides a predictive behavior by the statistics assessed with the available channels and their characteristics.

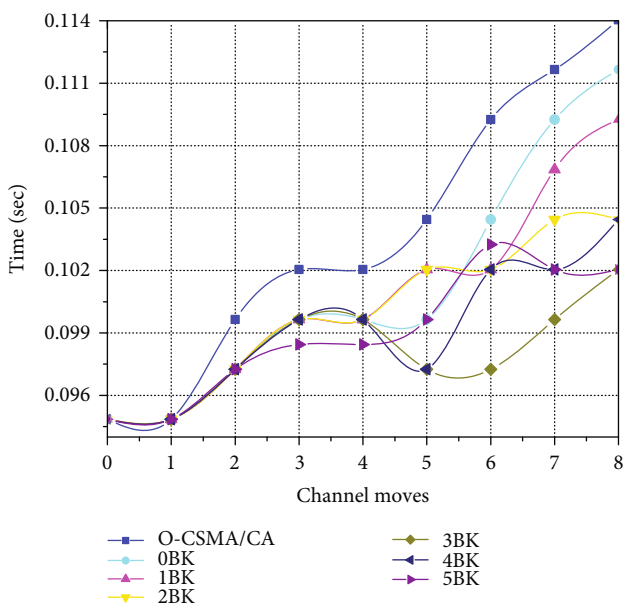


Figure 15: Coded-MAC vs. O-CSMA/CA without PU presence.

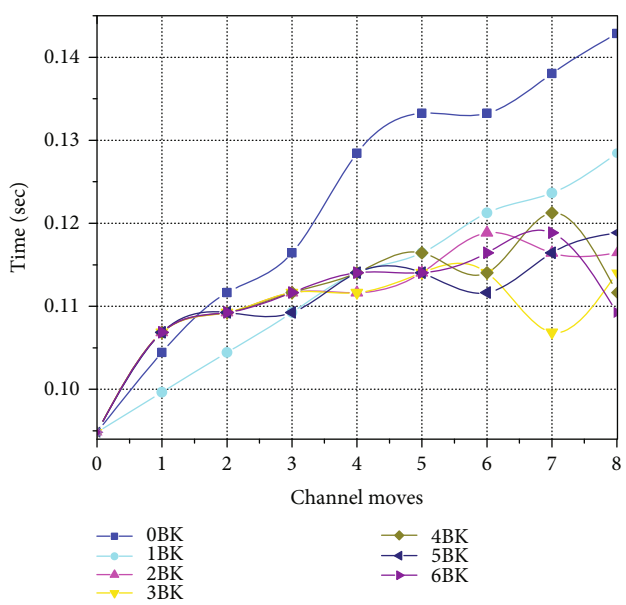


Figure 16: Coded-MAC for different BK channels to communicate with PU presence.

6. Conclusions

The proposal presented in this work demonstrates a decision-making module (DMM) which contributes to the functionalities required for opportunistic spectrum access (OSA). This function is based on coding for mitigation of the effects of frequent channel hops in the cognitive radio environment. The novel system solves spectrum allocation with simple resources and presents a dispersal algorithm that optimizes dynamically the spectrum allocation, while it is combined with the proposed MAC mechanism for the decision-making function recommended for cognitive radio networks.

The Reed Solomon (RS) code is used for establishing the communication process assisting medium access control (MAC). The communication code assigned by the base station (BS) identifies the device as a cognitive radio. This RS cognitive code (RSCC) is used to identify the device at the location and power range of the BS and identifies only cognitive device communications. The RSCC helps to gain fair access to the network for the cognitive radio (CR) devices because this registration process distinguishes the CR device from other opportunistic users in the area, and as a consequence, a more effective MAC can be implemented.

The use of an additional approach to the RS code was proposed using the information dispersal algorithm (IDA) which provides an effective redundancy to mitigate the effect of channel hops on CR networks (CRNs). It is important to diversify the information in order to optimize the spectrum use in cognitive radio networks, and the dispersal algorithm proposed is intended to dynamically allocate the communication in different available channels. This combined with the decision-making process to adapt spectrum utilization using intelligence by means of the multiple criteria decision-making module proposed.

The performance of the proposed system was proven when the backup (BK) channel availability was not always present, since BK channels can be occupied by a primary user (PU) at any unpredictable moment simulated by a probability uniform distribution. The proposed Coded-MAC performance was also tested for different packet sizes in order to find the most favorable ratio between channel hops and retransmissions of unacknowledged packets.

Data Availability

We already included the analysis data in our manuscript.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This project has been supported by grants awarded by the PRODEP program of the Mexican Secretariat of Public Education (SEP) and the National Council of Science and Technology (CONACyT).

References

[1] N. Chandwani, A. Jain, and P. D. Vyavahare, "Throughput comparison for cognitive radio network under various conditions of primary user and channel noise signals," in *2015 Radio and Antenna Days of the Indian Ocean (RADIO)*, pp. 1-2, Belle Mare, Mauritius, 2015.

[2] A. Naeem, M. H. Rehmani, Y. Saleem, I. Rashid, and N. Crespi, "Network coding in cognitive radio networks: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1945–1973, 2017.

[3] S. N. Kirillov and A. A. Lisnichuk, "Analysis of narrow-band interference effect on cognitive radio systems based on synthesized four-position radio signals," in *2018 XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE),*, pp. 50–54, Novosibirsk, Russia, 2018.

[4] N. Varshney, P. K. Sharma, and M. S. Alouini, "Opportunistic scheduling in underlay cognitive radio based systems: user selection probability analysis," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 191–195, Anaheim, CA, USA, USA, 2018.

[5] M. H. Rehmani and R. Dhaou, *Cognitive Radio, Mobile Communications and Wireless Networks*, Springer, Midtown Manhattan, NY, USA, 2019.

[6] X. Liu, M. Jia, W. Lu, F. Li, and D. Zou, "Cooperative spectrum sensing-based wideband cognitive radio system design," in *International Conference in Communications, Signal Processing, and Systems*, pp. 533–541, Springer, Singapore, 2017.

[7] B. Vishnu and M. A. Bhagyaveni, "Energy Efficient Cognitive Radio Sensor Networks with Team-Based Hybrid Sensing," *Wireless Personal Communications*, vol. 111, no. 2, pp. 929–945, 2020.

[8] I. Bajaj and Y. H. Lee, "Outage-constrained sensing threshold design for decentralized decision-making in cognitive radio networks," *IEEE Transactions on Communications*, vol. 64, no. 12, pp. 4956–4965, 2016.

[9] A. Sultana, X. Fernando, and L. Zhao, "An overview of medium access control strategies for opportunistic spectrum access in cognitive radio networks," *Peer-To-Peer Networking and Applications*, vol. 10, no. 5, pp. 1113–1141, 2017.

[10] A. Wyglinski, M. Nekovee, and Y. Hou, *Cognitive Radio Communications and Networks: Principles and Practice*, Academic Press, The Boulevard, Langford Lane, Kidlington, Oxford, UK, 2009.

[11] S. B. Mafra, R. D. Souza, J. L. Rebelatto, G. Brante, and O. K. Rayel, "Outage performance of a network coding aided multi-user cooperative secondary network," *Transactions on Emerging Telecommunications Technologies*, vol. 28, no. 2, article e2943, 2017.

[12] L. Pham, *Joint Source-Channel Coding for Image Transmission over Underlay Multichannel Cognitive Radio Networks*, 2019.

[13] Z. Zhao, Z. Ding, M. Peng, W. Wang, and J. S. Thompson, "On the design of cognitive-radio-inspired asymmetric network coding transmissions in MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 3, pp. 1014–1025, 2015.

[14] Y. Qu, C. Dong, D. Niu, H. Wang, and C. Tian, "A two-dimensional multiarmed bandit approach to secondary users with network coding in cognitive radio networks," *Mathematical Problems in Engineering*, vol. 2015, 10 pages, 2015.

[15] Y. Qu, C. Dong, S. Tang et al., "Opportunistic network coding for secondary users in cognitive radio networks," *Ad Hoc Networks*, vol. 56, pp. 186–201, 2017.

[16] W. Liang, H. V. Nguyen, S. X. Ng, and L. Hanzo, "Adaptive-ttcm-aided near-instantaneously adaptive dynamic network coding for cooperative cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1314–1325, 2015.

[17] Y. Qin, X. Zhong, Y. Yang, L. Li, and F. Wu, "Tcpjgnc: a transport control protocol based on network coding for multi-hop cognitive radio networks," *Computer Communications*, vol. 79, pp. 9–21, 2016.

[18] S. Khosroazad, A. Abedi, and N. Neda, "Achieving maximum bit rate in a cognitive radio network with physical layer network coding," *International Journal of Communication Systems*, vol. 31, no. 10, article e3558, 2018.

[19] P. Zhu, J. Li, D. Wang, and X. You, "Machine-learning-based opportunistic spectrum access in cognitive radio networks," *IEEE Wireless Communications*, vol. 27, no. 1, pp. 38–44, 2020.

[20] A. R. Sangi, M. S. Alkatheiri, S. Anamalamudi, and J. Liu, "Cognitive aodv routing protocol with novel channel-route failure detection," *Multimedia Tools and Applications*, vol. 79, pp. 8951–8968, 2020.

[21] S. Deng, B. Cao, X. Xiao, H. Qin, and B. Yang, "Cognitive routing optimization protocol based on multiple channels in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 16, no. 4, Article ID 155014772091451, 2020.

[22] S. Mei, B. Chen, F. Hu, and Z. Ma, "Hybrid network coding scheme in cognitive radio networks with multiple secondary users," *IEEE Access*, vol. 6, pp. 63948–63957, 2018.

[23] L. Sboui, Z. Rezki, and M.-S. Alouini, "Achievable rates of cognitive radio networks using multilayer coding with limited csi," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 395–405, 2016.

[24] R. El-Bardan, E. Masazade, O. Ozdemir, Y. S. Han, and P. K. Varshney, "Permutation trellis coded multi-level fsk signaling to mitigate primary user interference in cognitive radio networks," *IEEE Transactions on Communications*, vol. 64, no. 1, pp. 104–116, 2015.

[25] D. Hamza, K.-H. Park, M.-S. Alouini, and S. Aissa, "Throughput maximization for cognitive radio networks using active cooperation and superposition coding," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3322–3336, 2015.

[26] V. S. S. Nadendla, Y. S. Han, and P. K. Varshney, "Information-dispersal games for security in cognitive-radio networks," in *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1600–1604, Hong Kong, China, 2015.

[27] A. R. de Araujo Zanella and L. C. P. Albini, "A reed-solomon based method to improve message delivery in delay tolerant networks," *International Journal of Wireless Information Networks*, vol. 24, no. 4, pp. 444–453, 2017.

[28] B. Kumar, S. Kumar Dhurandher, and I. Woungang, "A survey of overlay and underlay paradigms in cognitive radio networks," *International Journal of Communication Systems*, vol. 31, no. 2, article e3443, 2018.

[29] V. Rajpoot and V. S. Tripathi, "A novel proactive handoff scheme with cr receiver based target channel selection for cognitive radio network," *Physical Communication*, vol. 36, p. 100810, 2019.

[30] C. Hernandez, C. Salgado, H. Lopez, and E. Rodriguez-Colina, "Multivariable algorithm for dynamic channel selection in cognitive radio networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, 2015.

[31] R. Rajmohan, K. Vishvaksenan, M. Mira, and S. Subramanian, "Performance of a turbo-coded downlink idma system using transmitter pre-processing," *Computers and Electrical Engineering*, vol. 53, pp. 385–393, 2016.

[32] X. Wu, "Embedded physical-layer authentication in cognitive radio requires efficient low-rate channel coding schemes," *IET Communications*, vol. 11, no. 3, pp. 400–404, 2017.

[33] Y. Xu, D. Li, Z. Wang, Q. Guo, and W. Xiang, "A deep learning method based on convolutional neural network for automatic

modulation classification of wireless signals," *Wireless Networks*, vol. 25, pp. 3735–3746, 2018.

[34] A. Lenz, A. Wachter-Zeh, and E. Yaakobi, "Duplication-correcting codes," *Designs, Codes and Cryptography*, vol. 87, no. 2-3, pp. 277–298, 2019.

[35] M. O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," *Journal of the ACM*, vol. 36, no. 2, pp. 335–348, 1989.

[36] M. Q. Naquid, R. M. Jimenez, and J. L. G. Compeán, "The babel file system," in *2014 IEEE International Congress on Big Data*, pp. 234–241, Anchorage, AK, USA, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8909458, 7 pages
https://doi.org/10.1155/2020/8909458

Research Article

Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention

Yan Chu,[1] Xiao Yue,[2] Lei Yu,[1] Mikhailov Sergei,[1] and Zhengkui Wang[3]

[1]*Harbin Engineering University, Harbin 150001, China*
[2]*Zhongnan University of Economics and Law, Wuhan 430073, China*
[3]*Singapore Institute of Technology, Singapore 138683*

Correspondence should be addressed to Xiao Yue; yuexiao@zuel.edu.cn

Received 18 January 2020; Accepted 24 September 2020; Published 21 October 2020

Academic Editor: Yin Zhang

Captioning the images with proper descriptions automatically has become an interesting and challenging problem. In this paper, we present one joint model AICRL, which is able to conduct the automatic image captioning based on ResNet50 and LSTM with soft attention. AICRL consists of one encoder and one decoder. The encoder adopts ResNet50 based on the convolutional neural network, which creates an extensive representation of the given image by embedding it into a fixed length vector. The decoder is designed with LSTM, a recurrent neural network and a soft attention mechanism, to selectively focus the attention over certain parts of an image to predict the next sentence. We have trained AICRL over a big dataset MS COCO 2014 to maximize the likelihood of the target description sentence given the training images and evaluated it in various metrics like BLEU, METEROR, and CIDEr. Our experimental results indicate that AICRL is effective in generating captions for the images.

1. Introduction

With the rapid development of digitalization, there are a huge amount of images, accompanied with a lot of related texts [1]. Automatic image captioning has recently attracted much research interest. The objective of automatic image captioning is to generate properly formed English sentences to describe the content of an image automatically, which is of great impact in various domains such as virtual assistants, image indexing, recommendation in editing applications, and the help of the disabled [2, 3]. Although it is an easy task for a human to describe an image, it becomes very difficult for a machine to perform such a task [4]. Image captioning does not only need to detect the objects contained in an image but also capture how these objects related to each other and their attributes as well as the activities involved in. Moreover, the semantic knowledge should be expressed in a natural language, which requires a language model to be developed based on the visual understanding.

Much research effort has been devoted to automatic image captioning, and it can be categorized into template-based image captioning, retrieval-based image captioning, and novel image caption generation [5]. Template-based image captioning first detects the objects/attributes/actions and then fills the blanks slots in a fixed template [1]. Retrieval-based approaches first find the visually similar images with their captions from the training dataset, and then the image caption is selected from similar images with captions [6]. These methods are able to generate syntactically correct captions but are unable to generate image-specific and semantically correct captions. Differently, the novel image caption generation approaches are to analyze the visual content of the image and then to generate image captions from the visual content using a language model [7]. Compared to the first two categories, novel caption generation can generate new captions for a given image that are semantically more accurate than previous approaches. Most of the works in this category rely on machine learning and deep learning, which is also the approach adopted in this paper. One common framework used in this category is the encoder-decoder framework for image captioning [8]. This framework was first introduced to describe a multimodal

log-bilinear model for image captioning with a fixed context window by Kiros et al. [9]. Recent research works have used the deep convolutional neural network (CNN) as the encoder and the deep recurrent neural network (RNN) as the decoder, which is proven to be promising [8, 10, 11]. However, it still remains challenging to identify the proper CNN and RNN models for the image captioning.

In this paper, we investigate one single-joint mode, AICRL, for automatic image generation using ResNet50 (a convolutional neural network) and LSTM (long short-term memory) with soft attention mechanism. AICRL consists of an encoder and a decoder. We adopt ResNet50 as the encoder to create an extensive representation of an input image by embedding it into a vector. Meanwhile, we utilize the LSTM with a soft attention as the decoder which selectively focuses the attention over a certain part of an image to predict the next sentences. Furthermore, we conduct extensive experiment and empirically determine the structure of the model and fine-tuned the model hyperparameters. Our experimental evaluation indicates that AICRL is effective to generate proper captions for the images.

The rest of the paper is organized as follows. Section 2 introduces the related work. In Section 3, we present the proposed AICRL model. Section 4 and Section 5 provide the experimental evaluation and conclusion, respectively.

2. Related Work

Much research has been devoted on the automatic image captioning recently. The research can be briefly categorized into three different categories including the template-based approaches, retrieval-based approaches, and novel image caption generation approaches.

The template-based approach is aimed at generating captions by using fixed templates with a number of blank slots, in which way different objects, attributes, and actions are detected first and then the blank spaces in the templates are filled. For example, Farhadi et al. [1] use a triplet of scene elements to fill the template slots for generating image captions. Li et al. [12] extract the phrases related to detected objects, attributes, and their relationships for this purpose. Kulkarni et al. [13] adopt a conditional random field (CRF) method to infer the objects, attributes, and prepositions before filling in the gaps. Template-based methods can generate grammatically correct captions. However, templates are predefined and length of captions cannot be variable.

The retrieval-based approach tries to generate description for an image by selecting the most semantically similar sentences from sentence pool or directly copying sentences from other visually similar images. For example, Gong et al. [6] utilize stacked auxiliary embedding method to generate image descriptions from millions of weakly annotated images. Ordonez et al. [14] find similar images in the Flickr database and return the descriptions of these retrieved images to query based on millions of images and their corresponding descriptions. Sun et al. [15] use semantic similarity and visual similarity scores to cluster similar terms and images together first and then retrieve caption of target image from captions of similar images in the same cluster. Hodosh et al. [16] establish a ranking-based framework to treat sentence-based image description as the task of ranking a set of captions for each test image. These methods generate general and syntactically correct captions. However, it is difficult for them to generate image-specific and semantically correct captions.

Different from the mentioned two categories, novel caption generation approaches mainly use deep learning and machine learning to generate the new captions. A general implementation of this method is to analyze the visual content of the image first and then generate image captions from the visual content using a language model. For instance, Vinyals et al. use CNN as an encoder for image classification and LSTM as a decoder to generate sentence for the description [8]. The main drawbacks of the work are the quick model overfitting, so they use the heavy and expensive GoogLeNet with 22 hidden layers and the absence of attention layer that significantly improved the description accuracy. Karpathy et al. investigate the possibility of generating an image description in natural language [10]. Their approach uses image datasets and their description in natural language and seeks an intermodal correspondence between words from the description and visual data. The first model aligns the fragments of sentences to the visual areas, then forms a single description by multimodal embedding. This description is treated as learning data for a second model of a recurrent neural network that learned to a generate caption. Xu et al. use a convolutional neural network to extract feature maps and LSTM to describe the input image, by processing already extracted feature maps [11]. The limitation of this work is the using of obsolete and expensive Oxford VGGnet, where the quality of image classification is low in the modern CNN [7]. Some researchers have put their attention on classification as Yu et al. [17], who propose a SVM classification-based two-side cross-domain algorithm by inferring intrinsic user and item features (CTSIF-SVMs), a two-side cross-domain algorithm with expanding user and item features via the latent factor space of auxiliary domains (TSEUIF) [18].

3. Model

In this section, we present our proposed model, AICRL, for automatic image captioning based on ResNet50 and LSTM with software attention. The ultimate purpose of AICRL is to generate the proper description for the given images. To do so, the AICRL model is designed with an encoder-decoder architecture based on CNN and RNN. In particular, to extract visual features, we use the ResNet50 network as the encoder to generate a one-dimensional vector representation of the input images. After that, to generate the description sentences, we adopt the LSTM as the language model for the decoder to decode the vector into a sentence. Meanwhile, we utilize the soft attention in the decoder to enable the model to selectively focus the attention over a certain part of an image to predict the next sentence better. We conduct extensive experiments, empirically determine the structure of the model, and fine-tune the model hyperparameters.

The whole model is fully trainable by using a stochastic gradient descent.

In the encoder-decoder method, the most likely description of the image is determined by maximizing the log-likelihood function of the expression S, considering the corresponding image I and the parameters of the model θ.

$$\theta^* = \arg \max_{\theta} \sum (I, S) \log p(S|I; \theta), \tag{1}$$

where θ is the parameter of our model, I is the input image, and S is the correct description. Since S represents a sentence of any length, therefore, a chain rule is usually used to model the joint probability over $S_1, \cdots, S_N$, where N is the length of this particular example.

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \cdots, S_{t-1}), \tag{2}$$

where the dependence on θ is omitted for convenience. The network training is represented by the pair of (S, I), and we optimize the sum of the log likelihood functions, as described in Equation (2), over the entire training set using stochastic gradient descent.

The likelihood $\log p(S_t|I, S_0, \cdots, S_{t-1})$ is modelled by a recurrent neural network, where there is a variable number of words that we define up to $t-1$. The hidden state of RNN (latent memory) h_t is updated after the new input x_t with the nonlinear function f.

$$h_{t+1} = f(h_t, x_t). \tag{3}$$

3.1. Image Feature Extraction. To represent the image, we adopt the convolutional neural network (CNN), ResNet50, which is a very deep network that has 50 layers. The depth of the network is crucial for neural networks, but deeper networks are more difficult to train. The structure of ResNet50 facilitates the training of networks and allows them to be much deeper, which leads to increased performance in different tasks. ResNet50 is much deeper than their "simple" counterparts, but moreover, the number of parameters (weights) of such networks is much smaller. For example, Table 1 indicates the number of parameter comparison between ResNet50 and VGG16. Deep convolutional neural networks have led to a series of breakthroughs for image classification. Recent evidence reveals that network depth is of crucial importance. Many other nontrivial visual recognition tasks have also greatly benefited from the deep models.

With the network depth increasing, the accuracy of networks increases rapidly, which is not surprising and then rapidly degrades (saturated). This degradation is not caused by overfitting, and the addition of even more layers leads to a higher learning error. In a sense, this is strange, since a deeper network has a strictly large representational power. It is possible for ResNet50 to get a deeper model trivially, which is not worse than the less deep network. It can be done by adding several identity layers, that is, levels that simply skip the signal further without changes. ResNet50's deeper levels have to predict the difference between the output of the previous layers and the objective function. They could always drive the weights to 0 and simply skip the signal. Hence, deep residual learning is a good method that makes the network learn to predict deviations from past layers.

Table 1: Comparison of total number of VGG16 and ResNet50 parameters.

CNN	Number of parameters
VGG16	138,357,544
ResNet50	23,587,712

The model takes an image and produces a caption, encoded as a sequence of $1-K$ coded words.

$$y = \{y_1, y_2, \cdots, y_c\}, y_i \in R^K, \tag{4}$$

where K is the size of the dictionary and c is the caption length. We use CNN in particular, ResNet50, to obtain set annotation vectors like the feature vectors. The extractor produces L-vectors, all of which is a D-dimensional representation of the corresponding part of an image.

3.2. The Language Model. The choice of f in Equation (3) is determined by its ability to cope with vanishing problems and exploding gradients, which are the most common problems in the design and training of RNN. LSTM networks are successfully used to accomplish the tasks of machine translation and sequence generation. In our design, we adopt LSTM as our language model to generate proper caption based on the input vector from the ResNet50 output.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{5}$$

where the output vector of the previous cell h_{t-1} with the new element of the sequence x_t is concatenated and passed as one vector through the layer with the sigmoid activation function.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}. \tag{6}$$

Two created vectors are used to update the state from C_{t-1} to C_t. To do this, we multiply the past state by f_t to "forget" the data recognized as unnecessary in the previous step, then add $i_t * \widetilde{C_t}$.

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\ \tilde{C}_t &= \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right). \end{aligned} \tag{7}$$

The input gate must determine what values will be updated, and the tanh layer creates a vector of new candidates for $\tilde{C}_t$, and values can be added to the cell state.

$$h_t = o_t * \tanh\left(C_t\right). \tag{8}$$

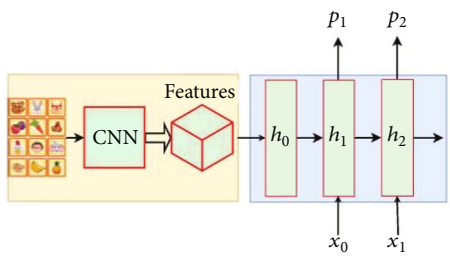


Figure 1: Model without attention.

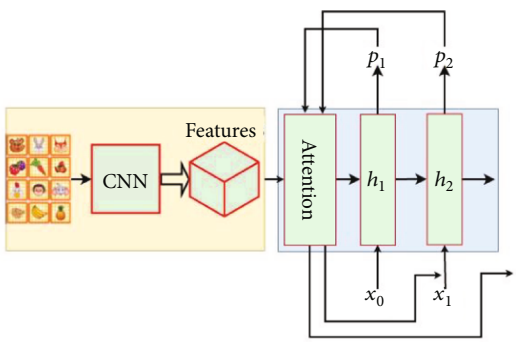


Figure 2: Model with attention.

The obtained values of C_t and h_t are transmitted to the neural network input at time $t + 1$.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),$$
$$p_t = \text{soft max}\ (h_t). \tag{9}$$

The multiplicative filters allow to effectively train LSTM, as they are good to prevent the exploding and vanishing gradients. Nonlinearity is provided by the sigmoid $\sigma(\cdot)$ and the hyperbolic tangent $h(\,)$. In the last equation, h_t is fed to the softmax function to calculate the probability distribution p_t over all words. This function is calculated and optimized on the entire training dataset. The word with maximum probability is selected at each time step and fed into next time step input to generate a full sentence.

3.3. Attention Mechanism. To better isolate the image content, we adopt the soft attention mechanism, which has been widely used to solve the problem of image classification, as there is no need to process all pixels of an image. For example, in the classification problem, the background usually plays an insignificant role. Nevertheless, convolutional neural networks, which are the most popular method for solving such a problem, spend the same amount of computational resources on all parts of the image.

Soft attention is implemented by adding an additional input of attention gate into LSTM that helps to concentrate selective attention. The main drawback of the model without attention is that it tries to decode the full image from the last hidden layer of h_0 in Figure 1. It is like an analogy with machine translation in the whole process. To do a translation of the whole text is just from the "last word." So it will lose a lot of useful information from the beginning of the text.

Table 2: Comparison for AICRL with and without attention.

Model	BLEU-4	METEOR	CIDEr
With attention	0.326	0.261	0.872
Without attention	0.262	0.209	0.803

Table 3: Comparison for AICRL with and without attention.

Model	Right choosing of generated description
With attention	71%
Without attention	54%

The attention gate can be represented as an addition input for LSTM in Figure 2. The soft attention depends on the previous output of LSTM p_t and extracted features of input image y_i. Soft attention is differentiable and can be trained by the standard method of the backpropagation algorithm. In the case of model with soft attention, we append an additional a_t in Equation (10).

$$a_t = \sum_{1}^{n} SjYj, \tag{10}$$

where a_t is an attention vector, s_j is a nonlinear function with softmax output, and y_j is the extracted features of the input image.

4. Experiments and Analysis

We perform an extensive set of experiments to evaluate the effectiveness of the proposed model. We have adopted two different datasets in our experiments including the MS COCO 2014 dataset and Flickr8K dataset, which contain the images with their descriptions in English. The MS COCO 2014 dataset contains 102,739 images with their descriptions, five descriptions for each image, and 20,548 testing examples. The Flickr8K dataset is another set of images with their descriptions with 7,000 training examples and 1,000 testing examples. Similar to MS COCO 2014, it also contains five descriptions for each image, but with a much smaller volume. Consider the Flickr8K data has less data than MS COCO 2014. In the training, we first use the Flickr8K dataset to train the model and then use the fine-tuned hyperparameters on MS COCO 2014. All experiments are conducted on NVIDIA GPU GTX-1070.

We evaluate the model using several popular metrics such as BLEU [19], METEOR [20], and CIDEr [21]. BLEU (Bilingual Evaluation Understudy) is an algorithm that measures the precision of an n-gram between the generated and reference captions. BLEU-N ($N = 1, 2, 3, 4$) scores can be calculated based on the length of the reference sentence, the generated sentence, the uniform weights, and the modified n-gram precisions.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an evaluation metric which was initially used in machine translation. Besides measuring precision,

Table 4: The performance comparison in the Flick8K dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Mao et al. [22]	0.58	0.28	0.23	—	—	—
Google NIC [28]	0.63	0.41	0.27	—	—	—
Chen and Zitnick [23]	—	—	—	0.141	—	—
Log bilinear [25]	0.656	0.424	0.277	0.177	0.173	—
DVS [26]	0.579	0.383	0.245	0.16	—	—
AICRL-ResNet50	*0.619*	*0.452*	*0.368*	*0.262*	*0.209*	*0.803*
AICRL-VGA16	0.672	0.436	0.338	0.225	0.186	0.743

Table 5: The performance comparison in the MS COCO 2014 dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Nearest neighbor [27]	0.48	0.281	0.166	0.1	0.157	0.383
Google NIC [28]	0.666	0.461	0.329	0.246	—	—
LRCN [24]	0.628	0.442	0.304	—	—	—
MS research [29]	—	—	—	0.211	0.207	—
Chen and Zitnick [23]	—	—	—	0.19	0.204	0.141
Log bilinear [25]	0.708	0.489	0.344	0.243	0.2	—
DVS [26]	0.625	0.45	0.321	0.23	0.195	0.66
AICRL-ResNet50	*0.731*	*0.562*	*0.41*	*0.326*	*0.261*	*0.872*
AICRL-VGA16	0.702	0.536	0.398	0.295	0.236	0.857

METEOR places emphasis on the recall between the generated and ground truth captions.

CIDEr (Consensus-based Image Description Evaluation) measures the similarity of generated captions to their ground truth sentences for evaluating image captioning. This measurement takes into account the grammaticality and correctness.

4.1. Training. The first step in the process of generating comments to the image is to create a fixed-length vector that effectively summarizes the content of an image. We use CNN, in particular the ResNet50 architecture. This network is preliminarily trained for 1.2 million images of the ImageNet dataset. Therefore, ResNet50 has a reliable initialization for object recognition and allows reducing training time. For any image from the training set, we get the output vector representation from the last convolution layer. This vector is fed to the LSTM input. Since the training set is a large dataset and each image is represented as a 2048-dimensional vector, the learning will be expensive. Therefore, the principal component method is used to reduce the dimension of the image vector from 2048 to 256. Since the length of the description may differ, the model should know where to start and stop. To do this, we add two tokens $< START >$ and $< END >$, which are the beginning and end of each sign.

The network for generating the captions will have to capture the words between these tokens. In this paper, words are represented as the frequency of occurrence of each word in the dictionary (1-of-N, where N is the power of the dictionary). The LSTM model learns to predict the next word S_t in the commentary based on the vector of visual features and the previous $t-1$ words. $p(S_t|I, S_1, S_2, \cdots, S_{(t-1)})$ is calculated and optimized on the whole training dataset by using stochastic gradient descent. At each time step, the context vector Z_t and the $h_{(t-1)}$ state of the previous step are fed to the LSTM together. After that, LSTM provides the next state vector h_t and next word. The context vector z_t is a concatenation of the feature vector and one hot vector of word representation.

4.2. Experimental Results. To speed up the learning process, we have adopted the method of Adam optimization with a gradual decreasing of learning rate which convergences more quickly. We use Adam optimization with regularization methods such as L_2 and dropout together. Applying the dropout technique in convolutional layers with a value of 0.5 and 0.3 in the LSTM layers helps to avoid overfitting that quickly happens with a small training set like the Flickr8K dataset. A variant with two LSTM layers is selected because we do not find that additional layers improve the quality. Each of the LSTM contains 512 hidden elements in a cell. Batch size equal to 32 and the beam size 3 are empirically found out that values are optimal. The deep models, such as ResNet50, for generating comments to the image increase in efficiency of the whole model. This is especially noticeable in the BLEU metric. Using a large set of MS COCO 2014 dataset avoids the model overfitting, while the overfitting on Flickr8K is achieved very quickly with a large batch size.

First, we study the impact of the soft attention mechanism in AICRL. As Table 2 indicates, the integrating the soft attention mechanism improves the model performances significantly. The soft attention mechanism increases the

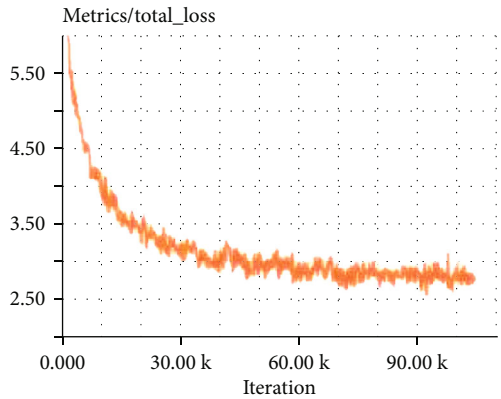


Figure 3: The total loss function.

performance in all metrics like BLEU-4, METEOR, and CIDEr. In addition, after training of the generator model, there are two questions. The first one is whether the model really generates new descriptions, and the second one is that whether they are diverse, qualitative, and understandable for humans. We have also conducted another set of experiments to involve human into the performance evaluation.

A questionnaire is designed with 20 images and the generated descriptions from the two different models. The participants are asked to evaluate whether the generated caption can well describe the images. Table 3 presents the results based on the generated description from the MS COCO 2014 dataset. From the results, we can see that 71% of the captions are well generated for the model with soft attention, while 54% are well generated for the one without soft attention. Based on this, we will use AICRL with the soft attention in the following experiments.

Next, we study the performance comparison between AICRL and other existing image captioning algorithms [22–29]. To make the evaluation complete, we also implemented another algorithm, AICRL-VGA16, by using another CNN network, namely, VGA16, in AICRL. Tables 4 and 5 show the results based on the Flick8K dataset and MS COCO 2014 dataset under six different metrics including BLEU-1, BLEU-2, BLEU-3, BLEU-4, METOER, and CIDEr. From both of the results, we can see that AICRL outperforms other systems in those metrics. The proposed model is able to generate efficient captions and fluent language. Meanwhile, ResNet50 also outperforms the VGA16 network which indicates that ResNet50 is able to capture the image features well. From these experiments, we observe that AICRL achieves good performance by integrating ResNet50, LSTM, and soft attention into a joint model.

Furthermore, we study how the total loss changes during the training. Figure 3 shows that the total loss of the model varies while the training iteration increases. From the results, we can see that the total loss quickly decreases at the beginning of training, but later, the speed of loss changing slows down.

5. Conclusions

In this paper, we have presented one single joint model for automatic image captioning based on ResNet50 and LSTM with software attention. The proposed model was designed with one encoder-decoder architecture. We adopted ResNet50, a convolutional neural network, as the encoder to encode an image into a compact representation as the graphical features. After that, a language model LSTM was selected as the decoder to generate the description sentence. Meanwhile, we integrated the soft attention model with LSTM such that the learning can be focused on a particular part of the image to improve the performance. The whole model is fully trainable by using the stochastic gradient descent that makes the training process easier. The experimental evaluations indicate that the proposed model is able to generate good captions for images automatically.

Data Availability

The data used to support this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61771155), China MOE Project of Humanities and Social Sciences for Youth (Grant No. 14YJC630181), the Natural Science Foundation of Hubei Province (Grant No. 2017CFB592), Fundamental Research Funds for the Central Universities Harbin Engineering University (Grant No. 3072020CF0608), and Fundamental Research Funds for the Central Universities Zhongnan University of Economics and Law (Grant No. 2722020JCT032 and No. 2722020PY047). This research was also supported in part by Singapore Ministry of Education TIF grant (Grant No. MOE2017-TIF-1-G018), Singapore Institute of Technology MOE Ignition grant (Grant No. R-MOE-E103-D004), and Strategic Initiative Grant on Applied Data Science.

References

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi et al., "Every picture tells a story: generating sentences from images," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., pp. 15–29, Springer, 2010.

[2] A. Graves, *Generating sequences with recurrent neural networks*, University of Toronto, 2013.

[3] A. Radford, L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, ICLR, 2016.

[4] Y. Zhang, R. Gravina, H. Lu, M. Villari, and G. Fortino, "PEA: Parallel electrocardiogram-based authentication for smart healthcare systems," *Journal of Network and Computer Applications*, vol. 117, pp. 10–16, 2018.

[5] M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, 2018.

[6] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image sentence embeddings using large weakly annotated photo collections," in *European Conference on Computer Vision*, pp. 529–545, Springer, 2014.

[7] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *Workshop on Neural Information Processing Systems (NIPS)*, 2014.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: a neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, Boston, MA, USA, 2015.

[9] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proceedings of the 31st international conference on machine learning (ICML-14)*, pp. 595–603, Beijing, China, 2014.

[10] A. Karpathy and L. Fei-Fei, *Deep visual-semantic alignments for generating image descriptions*, Stanford University, 2017.

[11] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *International conference on machine learning*, pp. 2048–2057, Lile, France, 2015.

[12] S. M. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. J. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics*, pp. 220–228, Portland, Oregon, USA, 2011.

[13] G. Kulkarni, V. Premraj, S. Dhar et al., "Baby talk: understanding and generating image descriptions," in *CVPR means IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2891–2903, 2011.

[14] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," *Advances in Neural Information Processing Systems*, pp. 1143–1151, 2011.

[15] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2596–2604, Santiago, Chile, 2015.

[16] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[17] X. Yu, Y. Chu, F. Jiang, Y. Guo, and D. Gong, "SVMs Classification based two-side cross domain Collaborative Filtering by inferring intrinsic user and item features," *Knowledge- Based Systems*, vol. 141, pp. 80–91, 2018.

[18] X. Yu, F. Jiang, J. Du, and D. Gong, "A cross-domain collaborative filtering algorithm with expanding user and item features via the latent factor space of auxiliary domains," *Patten Recognition*, vol. 94, pp. 96–109, 2019.

[19] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, 2002.

[20] S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translationand/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005.

[21] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, Boston, MA, USA, 2015.

[22] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, *Explain images with multimodal recurrent neural networks*, University of California, Los Angeles, 2014.

[23] X. Chen and C. L. Zitnick, *Learning a recurrent visual representation for image caption generation*, Stanford University, 2014.

[24] H. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched Long-Term Recurrent Convolutional Network for Facial Micro-Expression Recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 667–674, Xi'an, China, 2018.

[25] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *ICML '07: Proceedings of the 24th international conference on Machine learning*, pp. 641–648, Corvalis, OR, USA, 2007.

[26] A. Karpathy and L. Fei-Fei, *Deep visual-semantic alignments for generating image descriptions*, Stanford University, 2015.

[27] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[28] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: a neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2–7, Boston, MA, USA, 2015.

[29] H. Fang, S. Gupta, F. Iandola et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8822624, 12 pages
https://doi.org/10.1155/2020/8822624

Research Article

A Novel Ray-Casting Algorithm Using Dynamic Adaptive Sampling

Huadeng Wang,[1,2] Guang Xu,[1,2] Xipeng Pan,[1,2] Zhenbing Liu,[1,2] Rushi Lan,[1,2] and Xiaonan Luo[3]

[1]*School of Computer and Information Security, Guilin University of Electronic Technology, Guilin, China*
[2]*Guangxi Key Laboratory of Image and Graphic Intelligent Processing, China*
[3]*National Local Joint Engineering Research Center of Satellite Navigation and Location Service, China*

Correspondence should be addressed to Xipeng Pan; pxp201@guet.edu.cn

Received 23 May 2020; Revised 4 August 2020; Accepted 14 September 2020; Published 14 October 2020

Academic Editor: Yin Zhang

Ray-casting algorithm is an important volume rendering algorithm, which is widely used in medical image processing. Aiming to address the shortcomings of the current ray-casting algorithms in 3D reconstruction of medical images, such as slow rendering speed and low sampling efficiency, an improved algorithm based on dynamic adaptive sampling is proposed. By using the central difference gradient method, the corresponding sampling interval is obtained dynamically according to the different sampling points. Meanwhile, a new rendering operator is proposed based on the color value and opacity changes before and after the ray enters the volume element, and the resistance luminosity. Compared with the state of other algorithms, experimental results show that the method proposed in this paper has a faster rendering speed while ensuring the quality of the generated image.

1. Introduction

The ray-casting algorithm was proposed by Levoy in 1988. As an important volume rendering algorithm, the ray-casting algorithm is widely used in medical images, fluid and quantum mechanics, geographical exploration, and finite element analysis [1, 2]. In the basic principle of the ray-casting algorithm, each pixel in the imaging plane sends out a ray along the line of sight through the volume data, resampling along the ray at the same distance, obtaining the color value and opacity of each resampling point, and then synthesizing the color and opacity of each resampling point on the ray from the front to the back or from the back to the front.

Although the traditional ray projection algorithm has a high quality of synthetic image, it has two disadvantages. One is that there are a lot of volume data and a lot of addition and multiplication operations in the interpolation process, which makes the whole rendering process very slow. Second, after the sampling point is determined, the use of cubic linear interpolation will affect the real-time rendering process.

Some researchers use the closest interpolation instead of trilinear interpolation to remove some unnecessary sampling points. Some researchers achieve this goal by reducing the intersection of the light field and the data field. For example, Siddon [3] replaced the intersection of ray and voxel with the intersection of ray and plane. Ross et al. [4] employed the Bresenham algorithm to determine the intersection voxels. The data domain octree organization method proposed by Levoy [5] could easily skip empty voxels and reduce the time complexity of the intersection to O(nlog n). Jian [6] and others chose the projection light according to the correlation of the critical voxels, but the sharpness of the image would be worse.

There are two ways to improve the ray projection algorithm. One is to simplify the operation of the sampling process, such as using trilinear interpolation instead of complex cubic convolution interpolation [7–10]. The other is to

reduce the drawing time by reducing the number of sampling points on each ray. For the existing empty sampling points which are not helpful for image rendering, most of the methods are to eliminate them directly without considering the correlation between them and the surrounding voxels. When users interact with each other, the image will be blurred and the user experience will be affected.

Considering that most of the adaptive sampling methods used in most papers are equal-interval sampling, the threshold value of the isosurface is set first, and the isodistance sampling is started from the viewpoint direction along the line of sight direction until the sampling value exceeds the set threshold value. Finally, the linear interpolation is done between the sampling points $P(n)$ and $P(n+1)$ to get the intersection coordinates of the line of sight and the isosurface. The disadvantage of this method is that there are too many sampling points. Therefore, a large number of sampling points are empty voxels, which wastes precious time and is not conducive to user interaction. Another method is based on a given sampling frequency, but the quality of voxel rendering is very poor. Some scholars try to improve the sampling frequency, adjust the sampling frequency by the distance proportion between the tangent point of the surface normal and the viewpoint, and simplify the image synthesis operator. Another is to resample at a lower sampling frequency first. If the data values of two adjacent resample points differ greatly, then increase the sampling frequency in such a high-frequency area. In the place where the voxel value changes slowly and quickly, the sampling interval should be small to avoid missing voxel information. If the voxel value changes slowly, increase the sampling interval appropriately, so that a large number of empty voxels can be skipped. However, if the sampling interval is too large, the rendering quality will be low, resulting in a large "void" effect [11, 12], that is, the phenomenon of the artifact.

There are two shortcomings in the current ray casting algorithm for medical image rendering, one is the slow rendering speed, the other is the low efficiency and high time complexity of the whole sampling process. In order to solve the above shortcomings, this paper proposes an algorithm. The algorithm uses the central differential gradient method. By this method, the specific sampling distance can be determined dynamically according to the different distances of sampling points, which can improve the sampling efficiency. On the other hand, according to the change of color value and opacity before and after the ray enters the voxel, we propose a new rendering operator to improve the final image rendering speed.

This paper solves the problem of how to find the most suitable sampling interval dynamically. Through the full increment formula of the voxel value difference between two adjacent sampling points, combined with the central difference method, the calculated gradient, and the preset voxel value threshold, the sampling point interval can be calculated dynamically according to the different positions of sampling points. Thus, we can update the sampling interval dynamically on the premise of ensuring the sampling density. Aiming at the problem that the search precision of the intersection point between the light and the actual isosurface will be reduced when the sampling point interval is large, this paper proposes a recursive estimation method based on the intersection point to estimate the intersection point between the sampling point and the isosurface, which is simple and easy to operate. The algorithm we proposed is a novel ray-casting algorithm using dynamic adaptive sampling, which is called the ray-casting DAS algorithm for short. After determining the sampling point, aiming at the general paper's shortcomings of using the nearest interpolation or trilinear interpolation to synthesize the image according to the opacity and color value of eight data points equidistant from the sampling point, this paper adopts a method of light blocking, which is to change the opacity and color values as well as the photometry before and after the ray enters the volume element and improve the composition operator. The basic algorithm structure and process are listed in Figure 1. Experimental results show that the improved method can effectively improve the speed of image rendering and avoid the generation of artifacts. The main innovation of this paper is dynamic adaptive sampling and improving the rendering operator to improve the efficiency and rendering time of the algorithm. Another unique innovation of this paper is the introduction of the distance function to improve the image noise and rendering quality.

The remainder of this paper is as follows. The second part summarizes the latest research results of some ray projection algorithms and compares their advantages and disadvantages. In the third part, the ray projection algorithm based on adaptive sampling is introduced. The fourth part is the comparison between the ray-casting DAS and other ray projection algorithms in several open-source datasets. The fifth part is the summary of ray-casting DAS and the assumption of improving in the future.

2. Related Work

Wu et al. [13] proposed a GPU ray projection method for visualizing 3D pipelines in a virtual sphere. In their method, the pipeline data was initially divided into several data blocks. The pipeline centerline in each block was then segmented and coded, and a thicker pipeline envelope was then constructed using geometric shaders. Finally, elaborate 3D pipes could be rendered using pixel shaders. Compared with the traditional polygon-based pipeline data partition method, this method improves the rendering frame rate under the same pixel-level precision display effect. The visualization of 3D pipe thickness is realized without affecting the rendering efficiency.

Luo et al. [14] proposed a rendering algorithm based on the height field [15] and an improved algorithm of finding the intersection of isosurfaces to solve the problem of the slow speed of finding the intersection of light and volume data. However, because the volume data will be imaged many times due to the reflection of the surface of the object, so it needs to do many volume data rendering operations, which hurt the performance of the algorithm.

Francisco Sans et al. [16] proposed to combine CPU based ray projection with GPU, and reduce the time complexity of rendering by dynamic gradient interpolation of

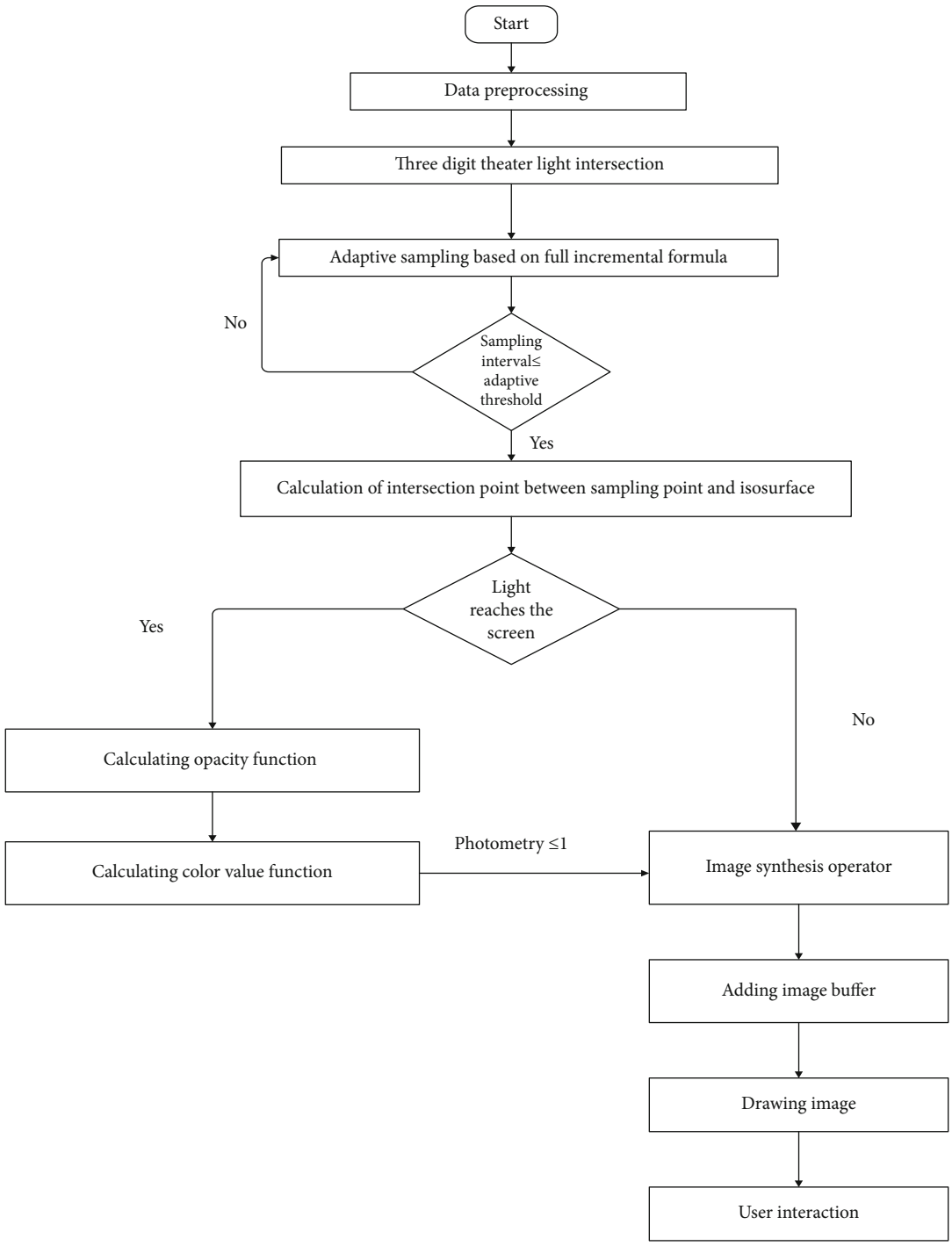


Figure 1: The brief structure of the ray-casting DAS algorithm.

voxels. It realized single channel GPU rendering and meet the real-time interaction requirements of users.

Binyahib et al. [17] proposed a parallel volume rendering algorithm which combines the classic object sequence and image sequence techniques. The algorithm runs on an unstructured grid (and structured grid), so it can deal with block boundary crossing in a complex way. It can also effectively deal with situations that are prone to load imbalance [18]. At the largest scale, it can be expanded to 8192 processors and run on datasets of more than 1 billion cells.

Tao et al. [19] proposed a ray projecting algorithm combining the fitness estimation based on ray-casting with the global optimization method of improved adaptive differential evolution. This method eliminated the fine registration step of the famous iterative nearest point (ICP) algorithm [15, 20, 21], and it is the first direct global registration algorithm.

Moreover, the calculation of the fitness of the global optimization method was accelerated, and the search space was effectively used to find the optimal transformation solution. The algorithm was successfully implemented in parallel mode on multicore computer processors.

3. Our Proposed Approach

3.1. Data Preprocessing Based on Fuzzy Enhancement. Because in the ray projection algorithm, in the mapping from three-dimensional to two-dimensional, the image edge, region, texture, and so on are inevitably blurred due to the loss of information. Therefore, the algorithm proposed in this paper first preprocesses the data, uses the method of enhancement operator [22] to calculate the blur rate of the enhanced image, and then selects the best threshold.

The purpose of introducing a fuzzy enhancement operator [23] is to enhance the contrast of the image region by different enhancement processing, that is, the black target area is attenuated and the white background area is enhanced. The enhanced image not only retains the information of the original image but also makes the layers of each region clearer, which further reduces the image ambiguity. In this way, the change of the processed data is beneficial to the extraction and selection of sampling points for the subsequent volume rendering algorithm.

3.2. Intersection Processes of Light and Plane. First, the parametric equation of light emitted by a pixel in the plane can be expressed as:

$$\begin{cases} x = x_0 + lt, \\ y = y_0 + mt, \\ z = z_0 + nt, \end{cases} \tag{1}$$

where (l, m, n) represents the direction vector of the line in the object space, (x_0, y_0, z_0)represents a point on the line. x, y, and z denote the three-dimensional coordinates of any point on the line, and t is the parameter of the linear parametric equation, indicating the number of directions of the line.

Let the size of the regular data field be $L \times M \times N$ and logical coordinates $(x, y, z) = (i \times X, j \times Y, k \times Z)$, in which, $i = 0, 1 \cdots, L-1$, $j = 0, 1, \cdots, M-1$, and $k = 0, 1, \cdots, N-1$. Here L, M, and N denote the number of X, Y, and Z plane families, respectively, and (x, y, z) represents the projection coordinates of a specific point in the X, Y, Z plane of the specific plane family. The data distribution of the volume data field is represented by the plane family. The definition of the plane family is as follows:

$$\begin{cases} x = i \times X, i = 0, 1, \cdots, L-1, \\ y = j \times Y, j = 0, 1, \cdots, M-1, \\ z = k \times Z, k = 0, 1, \cdots, N-1. \end{cases} \tag{2}$$

Because the normal vector of light $(l, m, n) \neq 0$, let us set $n \neq 0$. Then, the parameters of the intersection of light and plane $\{z = k \times Z, k = 0, 1, \cdots, N-1\}$ can be calculated as follows:

$$t_k = \frac{k \cdot Z - z_0}{n}, k = 0, 1, \cdots, N-1. \tag{3}$$

Further, the intersection of light and $z = k \times Z$ can be expressed as:

$$\left(x_0 + \frac{(-z_0 + kZ) \cdot l}{n}, y_0 + \frac{(-z_0 + kZ) \cdot m}{n}, kZ\right), k = 0, 1, \cdots, N-1. \tag{4}$$

Eq. (4) infers that light and two adjacent planes $z = k \times Z$ and $z = (k+1) \times Z$ intersection parameters t_k and t_{k+1} satisfying Eq. (5)

$$t_{k+1} = t_k + \frac{Z}{n}. \tag{5}$$

The intersection points of the straight line and z plane family can be obtained by using the above iteration formula:

$$\begin{cases} x_{k+1} = x_k + \dfrac{Z}{n}, \\ y_{k+1} = y_k + \dfrac{m}{n} Z, k = 0, 1, \cdots, N-2, \\ z_{k+1} = z_k + \dfrac{Z}{n}. \end{cases} \tag{6}$$

Similarly, the recurrence formula to the intersection of the line and the x, y plane can be obtained.

3.3. Dynamic Adaptive Sampling. The intersection points of the line and plane family obtained by the above operations are not the sampling points we need, but we can get the sampling points on the line based on these intersections. The general method is to obtain the sampling points by equidistant sampling with a fixed distance d. Here, we consider how to dynamically adjust the adaptive sampling interval.

Let g be a three-dimensional volume data field. According to the total differential formula of the binary function, the difference between the voxel values of two adjacent sampling points can be obtained as follows:

$$\begin{aligned} \Delta g &= g(n+1) - g(n) \\ &= g'_x \cdot \Delta x + g'_y \cdot \Delta y + g'_z \cdot \Delta z + o(\rho) \approx \nabla g \cdot L \cdot d_n, \end{aligned} \tag{7}$$

where ∇g is the gradient of the 3D digital theater and d_n is the coordinate distance of the sampling point. L is the unit vector of the light direction.

Because $g(n)$ has been estimated, and $g(n+1)$ is unknown. It may be assumed that the voxel value $g(n+1)$ of the sampling point $P(n+1)$ is T (T is the set voxel value threshold). According to Eq. (7), the sampling interval can be approximately expressed as:

$$d_n \approx \frac{T - g_n}{\nabla g \cdot L}. \tag{8}$$

```
Input: P(n), P(n + 1) number_Current, Total.
Output: P0
Take the coordinate P0 of the geometric midpoint of P(n) and P(n + 1) to obtain the sampling value g.
T is set threshold, and the number current represent the current iterations, which is required to less than the Total.
If g > T then
    P(n + 1) = P0
Else
    P(n) = P0
While |P(n) − P(n + 1)| < ε do
        By linear interpolation of P(n) and P(n + 1), the coordinates of the intersection point of P(n) and P(n + 1)
If number_Current>Total then return P0
  If P0 is close to P(n) then
      P(n) = P0
  Else
      P(n + 1) = P0
  Return P0
```

Algorithm 1. solution process of intersection coordinates

It can be seen from Eq. (8) that if the projection value of the gradient at a point in the direction of light is smaller, and the difference between $g(n)$ and T is large, then the interval d_n will be larger; otherwise, d_n will be too small. In order to prevent the sampling interval from being too large or too small, an upper and lower bound can be added to the sampling interval of Eq. (8). It is as follows:

$$d_n = \begin{cases} D_1 & 0 \le \dfrac{T - g_n}{\nabla g \cdot L} \le D_1, \\ D_2 & \dfrac{T - g_n}{\nabla g \cdot L} \ge D_2, \text{ or } \dfrac{T - g_n}{\nabla g \cdot L} < 0, \\ \dfrac{T - g_n}{\nabla g \cdot L} & D_1 < \dfrac{T - g_n}{\nabla g \cdot L} < D_2. \end{cases} \tag{9}$$

D_1 and D_2 are upper and lower bounds of sampling, respectively.

In order to simplify the calculation, the center difference method is used to calculate the gradient, and the calculation method is as follows:

$$\nabla g = \frac{1}{2} \begin{cases} g(x+1, y, z) - g(x-1, y, z), \\ g(x, y+1, z) - g(x, y-1, z), \\ g(x, y, z+1) - g(x, y, z-1). \end{cases} \tag{10}$$

Because the gradient calculation will affect the rendering speed if it is placed in the sampling stage, so we put the gradient calculation in the data preprocessing stage and use a linear table to store the calculated gradient value. When calculating the sampling step length, we look up the gradient value closest to the sampling point in the linear table as the gradient of the sampling point. Assuming the number of sampling points is n, then the time complexity of this step is $O(n)$.

After determining the sampling interval, we need to find the intersection of two adjacent sampling points and the isosurface and then synthesize the image through the color value and opacity of the intersection. In order to improve the imaging quality, this paper presents a recursive method for intersection estimation.

Supposing that we need to find the intersection point of the two sampling points $P(n)$ and $P(n+1)$ and the isosurface. First, we take the midpoint coordinate P0 of $P(n)$ and $P(n+1)$. If the corresponding voxel value at P0 is greater than the set threshold T, then P0 is taken as the new $P(n+1)$. If the corresponding voxel value at P0 is less than the set threshold T, then P0 is taken as the new $P(n)$, and then continue to take the midpoint of $P(n)$ and $P(n+1)$ until the distance between $P(n)$ and $P(n+1)$ is very small (less than the given value). Then, the final $P(n)$ and $P(n+1)$ are linearly interpolated to obtain the coordinates of P0, which is the coordinates of the intersection.

The process of finding the intersection coordinates is shown in Algorithm 1. Through the algorithm, we can calculate the intersection point of sampling points $P(n)$ and $P(n+1)$ fast. Moreover, we set the threshold of the number of iterations, which improves the convergence speed of the algorithm and indirectly improves the rendering speed of the image.

3.4. Improving Rendering Operator. After selecting the sampling point and intersection point in the above steps, the traditional method is to do linear interpolation around the eight closest data points of the intersection point to synthesize the image [24]. For example, the ray projection algorithm based on trilinear interpolation for fiber surface is adopted by Francisco Sans, and the time complexity of the trilinear interpolation is $O(n^3)$.

The formula of cubic linear interpolation is as follows:

$$\begin{aligned} S(\mathrm{i}, j, k) &= P_{000}(i-1)(j-1)(k-1) \\ &+ P_{100} i(j-1)(k-1) + P_{010}(i-1) j(k-1) \\ &+ P_{110} ij(k-1) + P_{001}(i-1)(j-1)k \\ &+ P_{101} i(j-1)k + P_{011}(i-1)jk + P_{111} ijk. \end{aligned} \tag{11}$$

Among them, $S(i, j, k)$ represents the voxel value of the corresponding spatial coordinate point $P(i, j, k)$, P000, P001, $\cdots$, P111 represents the voxel value of the eight adjacent data points of $P(i, j, k)$, respectively.

In this paper, we consider a method of ray blocking and synthesize a new rendering operator through the change of color and opacity before and after the ray enters and leaves the voxel, as well as the corresponding opacity. The experimental results show that it can effectively reduce the time complexity and improve the rendering speed without affecting the imaging quality.

Let the opacity and color values of the ith volume element be P_i and Q_i, respectively, the opacity and color values before the ray enters the ith volume element P'_i and Q'_i, respectively, and the opacity and color values after the ray enters are P''_i and Q''_I, respectively, and let the opacity of the cell be G_i; the composite operator of the point image is

$$P''_i \times Q''_i = P_i \times Q_i \times (1 - G_i) + P'_i \times Q'_i. \quad (12)$$

The relationship between the opacity and opacity of current voxels is

$$P''_i = P_i \times (1 - G_i) + P'_i. \quad (13)$$

At the same time, we notice that not all the light can finally reach the imaging screen for imaging. The initial point has the smallest opacity. With the increase of data points through which the ray passes, the opacity also increases. Finally, after the ray enters the nth data point, the opacity becomes the maximum 1, so the image synthesis operator can be simplified as:

$$Q''_i = \sum_{i=1}^{n} P_i \times Q_i \times (1 - G_i). \quad (14)$$

From the above formula, it can be found that when $g = 1$, that is the maximum photometry, the right side of the equation is equal to 0, which means that the ray that does not reach the screen does not participate in the calculation of the synthesis operator. At the same time, the time complexity of single P_i and Q_i calculation is $O(n)$, so the time complexity of the final synthesis operator is $O(n^2) < O(n^3)$, so the algorithm effectively improves the rendering efficiency. On the other hand, it can not only improve the time of image rendering but also effectively prevent the generation of artifacts and improve the quality of rendering.

4. Experimental Results

In this section, several experiments and comparison results are reported to evaluate the performance of the proposed algorithm ray-casting DAS. We compare the ray-casting DAS with other algorithms in several aspects, such as time complexity, image rendering speed, and image quality.

4.1. Time Consumption Comparison. The following is a comparison of several algorithms' test data on the open dataset ISIC2018 [25]. Table 1 compares the drawing time of different algorithms for drawing the same picture of the human leg. VS3D, Dual-FHR, RIFT2, SHSR-UV, and ray-casting DAS in Figure 2 represent the experimental results of Wu's algorithm [13], Luo's algorithm [14], Francisco Sans' algorithm [16], Roba Binyahib's algorithm [17], and ray-casting DAS algorithm, respectively.

Table 1: Comparison of the time consumption of each stage of the above algorithms (unit: seconds).

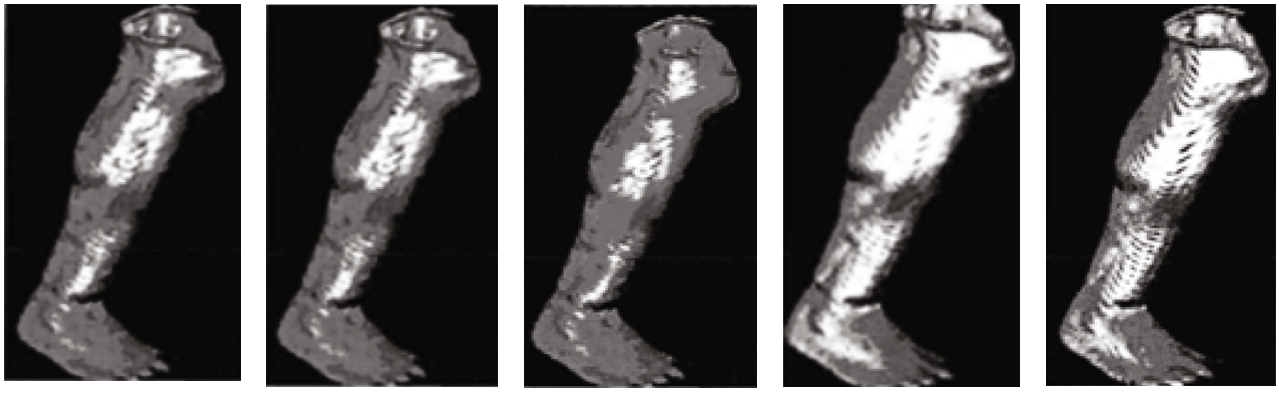
Algorithms	Data preprocessing time	Sampling time	Drawing time	Total time
VS3D	23	32	33	88
Dual-FHR	17	28	31	76
RIFT2	16	25	31	72
SHSR-UV	17	23	26	66
Ray-casting DAS	24	16	21	61

According to the experimental results in Table 1, compared with other algorithms, the preprocessing time of the algorithm in this paper is increased by about 6 to 7 seconds compared with other algorithms. This is because we add gradient calculations in the preprocessing stage, which consumes more time. However, due to the use of dynamic adaptive sampling, the acquisition of sampling points is more efficient, and the interference of unnecessary sampling points is eliminated. Compared with other algorithms, the time consumption of this algorithm in the sampling phase is greatly reduced, reducing 7-16 seconds. On the other hand, due to the improvement of the rendering operator, the idea of early termination is added in the rendering process, which greatly reduces the rendering time by 5-10 seconds. Finally, compared with other better algorithms, the total time is improved by 5-20 seconds, and the speed is significantly improved.

Through the results of Figure 2, we can find that the algorithm in this paper can better draw the details of the leg, including some texture information. These details are very valuable for medical diagnosis. Compared with other algorithms, the overall occlusion of the scene is captured coherently and close to the real value. The shadow area of the leg is less, and more texture details are preserved with ray-casting DAS.

4.2. Rendering Speed and Quality Comparison on CT Dataset. As shown in Figure 3, RIFT2, SHSR-UV, ICVC-GPU, DOS-Ray, IMPA, and ray-casting DAS represent the experimental results of Francisco Sans' algorithm [16], Roba Binyahibs' algorithm [17], Feiniu Yuan's algorithm [26], Leonardo's algorithm [27], Yan Zhang's algorithm [28], and ray-casting DAS algorithm, respectively. By comparing the results of the different algorithms on the dataset of ISBI2018 [29], the ray-casting DAS algorithm can reduce and avoid the phenomenon of artifact in CT image rendering and improve the visibility of the image.

Table 2 compares the rendering time of the ray-casting DAS algorithm with the other five algorithms. The



(a) VS3D (b) Dual-FHR (c) RIFT2 (d) SHSR-UV (e) Ray-casting DAS

Figure 2: The experimental results of the above algorithms on dataset ISIC2018. The experimental results show that compared with other algorithms, this algorithm can significantly reduce the sampling time required for image rendering, improve the rendering speed, and also improve the overall time consumption.

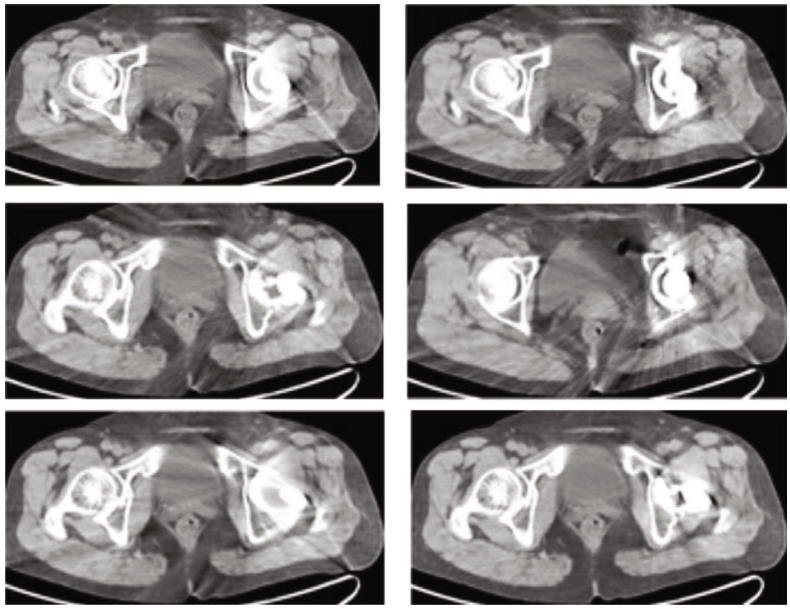


Figure 3: Comparison of the phenomenon of artifact and respective visibility of several algorithms.

Table 2: Comparison of the rendering speed of various algorithms (drawing times/s, signal-to-noise ratio/dB).

Algorithms	Iteration	Drawing time	Signal-to-ratio
RIFT2 algorithm [16]	46	24	38
SHSR-UV algorithm [17]	53	29	42
ICVC-GPU algorithm [26]	34	21	33
DOS-ray algorithm [27]	49	25	34
IMPA algorithm [28]	45	22	36
Ray-casting DAS algorithm	32	15	60

experimental results show that ray-casting DAS has better performance in rendering time and better image signal-to-noise ratio than the previous algorithms.

Similarly, from the experimental results in Figure 3, it can be seen that the rendering algorithm in this paper can effectively solve the problem of artifacts in CT image rendering. From (a) to (e), we can clearly see the existence of artifacts in CT images, they obviously cause interference to the analysis image, and the image (f) drawn by ray-casting DAS algorithm can effectively eliminate the existence of these artifacts. It can be clearly seen from the graph that in the experimental results of this paper, each part of the image has no artifacts interference, and the image is clear.

4.3. Influence of Different Distance Functions on Image Rendering. The distance function expresses the relationship between distance and distance intensity value. The distance determines the distance intensity value. Through the distance function, the distance and depth are reflected in the final drawing process. This paper proposes an interactive

Table 3: Comparison of several distance functions.

Algorithms	LCC	SPOCC	SSIM	MSSIM	PSNR
Without distance function	0.69	0.71	0.792	0.813	29 db
Ray-casting DAS with f1 function	0.72	0.76	0.814	0.839	33 db
Ray-casting DAS with f2 function	0.76	0.79	0.857	0.874	36 db
Ray-casting DAS with f3 function	0.75	0.78	0.849	0.867	35 db
Ray-casting DAS with f4 function	0.86	0.85	0.896	0.918	43 db

algorithm. Interaction is based on distance first. The interaction process specifies the distance value, and the distance value is used in the distance function. Then, the final transparent value is determined by combining the distance function in the rendering process. The corresponding distance function is a piecewise function:

$$f(x) = \begin{cases} \text{high}, x \in [0, k), \\ \text{low}, x \in [k, \max). \end{cases} \tag{15}$$

The value k is the distance value obtained by interaction. When the distance is less than k, it is a high-intensity value, and when the depth is greater than k, it is a low-intensity value.

In order to improve the quality of rendering and reduce the interference of noise and other factors, we introduce the concept of distance function before color and opacity composition. The four contrast distance functions used in this experiment are as follows:

$$f_1(x) = \max_i \{\|x_i\|, 1 \le i \le p, \|x_i\| \le 52\}, \min f_1 = 0, \tag{16}$$

$$f_2(x) = \sum_{i=1}^{p} i\|x_i\|^4 + \text{random}[0, 1), 1 \le i \le p, \|x_i\| \le 52, \min f_2 = 0, \tag{17}$$

$$f_3(x) = \sum_{i=1}^{p} \|x_i\|^2 - 10 \cos(2\pi\|x_i\|), 1 \le i \le p, \|x_i\| \le 52, \min f_3 = 0, \tag{18}$$

$$f_4(x) = -20 \sum_{i=1}^{p} \exp\left(-0.2\sqrt{\frac{1}{30}\|x_i\|^2}\right) - \frac{1}{30} \exp\left(\sum_{i=1}^{p} \cos(2\pi\|x_i\|)\right) + 20 + e, 1 \le i \le p, \|x_i\| \le 52, \min f_4 = 0. \tag{19}$$

Here, x_i is a three-dimensional column vector, representing the spatial coordinates of the ith voxel, and p is the total number of all effective voxels.

As is shown in Table 3, LCC [31] is short for linear correlation coefficient, and SROCC [32] is short for Spearman's rank-order correlation coefficient, which are both evaluation indexes to measure the similarity between drawn image and real image. The structural similarity theory is simplified as SSIM [33], whose value is between 0 and 1. The larger LCC, SROCC, and SSIM value indicate the higher similarity between the drawn image and the real value of the original image. The peak signal-to-noise ratio (PSNR) [32] measures the ratio of useful information and noise in an image. For the same image, the larger the signal-to-noise ratio is, the smaller the noise is, and the higher the image quality is.

In the experiment, four different distance functions are used, and their performances are compared. The experimental results in Table 3 show that, compared with the method without distance function, adding distance function can significantly increase the similarity between the drawn image and the real image; at the same time, it has a higher signal-to-noise ratio and relatively less noise content. On the other hand, Figure 4 shows that different distance functions have advantages and disadvantages in image rendering. The distance function f_4 is the best in the quality of completion. It can eliminate the interference of noise to the greatest extent. The image drawn is also the clearest and contains the least noise interference.

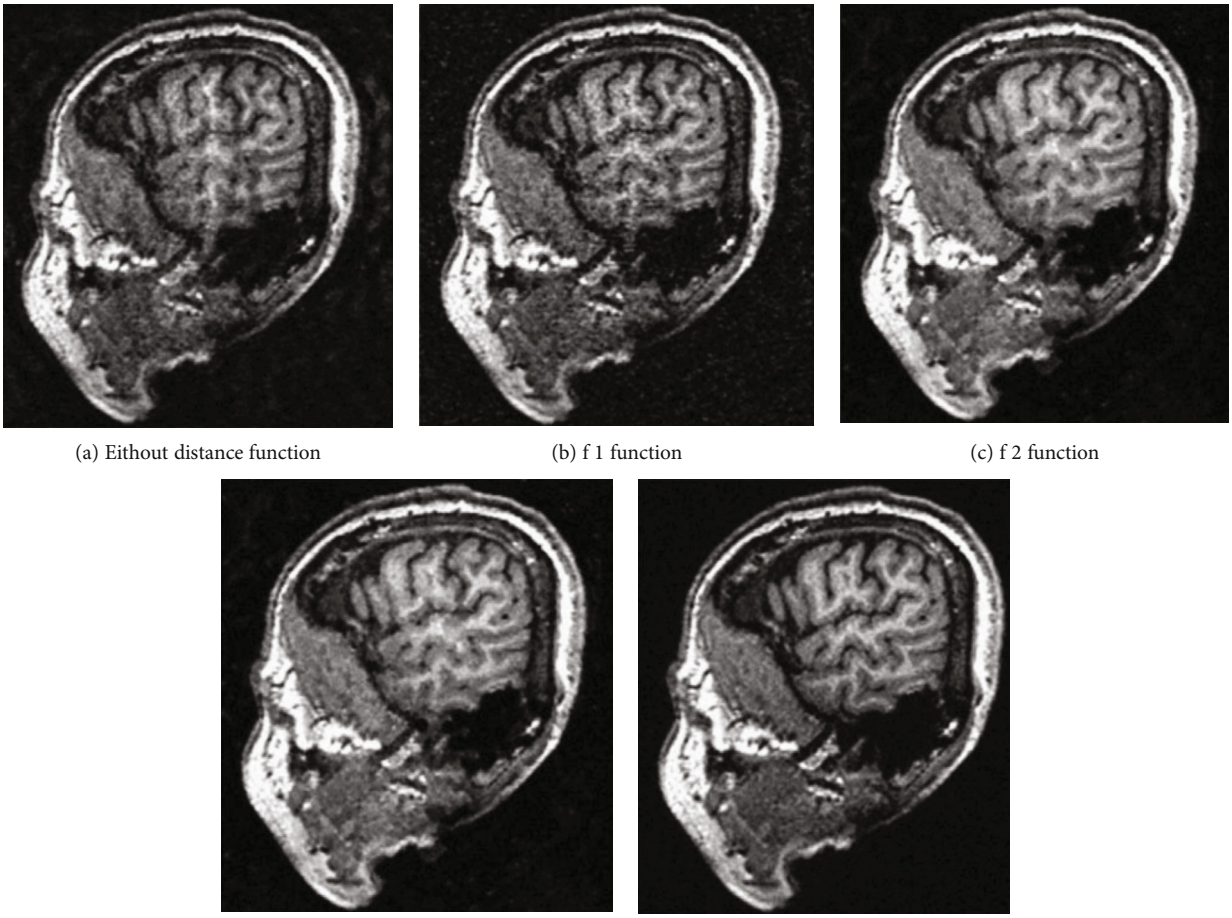
4.4. Image Rendering Quality Comparison on Other Dataset. We use the open-source dataset VisMale [34] for experiments. The simulation environment is VS2016 and OpenGL. Table 4 shows the comparison of the six algorithms for human head image rendering, and ICVC-GPU, MVRC2, SigIg+FMI, SHSR-UV, DBRay, and Ray-casting DAS in Figure 5 represent the experimental results of Feiniu Yuan's algorithm [26], Mohammandmehdi Bozorgi's algorithm [32], R. Mehaboobathunnisa's algorithm [35], Roba Binyahibs' algorithm [17], Alec G.Moore's algorithm [36], and ray-casting DAS algorithm proposed in this paper, respectively.

In Figure 6, information entropy, contrast, sharpness, SNR [31, 33], and PSNR [32] are three objective indicators reflecting image quality. Their definitions are as follows:

(1) Information entropy represents the amount of information contained in the image, which is proportional to the amount of information contained in the image [33]. The formula is

$$\text{Ent} = -\sum_{i=0}^{L-1} \sum^{L-1} p(i, \hat{j}) \log p(\hat{i}, j), \tag{20}$$

where i represents the gray value of the pixel $(0 \le i \le 255)$, j represents the mean gray value of the domain $(0 \le j \le 255)$. And the formula $p(i, j) = f(i, j)/\text{N2}$ reflects the comprehensive characteristics of the gray value of a pixel position and the gray distribution of its surrounding pixels, where $f(i, j)$ is the frequency of the feature binary (i, j), and N is the scale of the image.



(a) Eithout distance function (b) f 1 function (c) f 2 function

(d) f 3 function (e) f 4 function

Figure 4: Comparison of algorithms using different distance functions in the MICCAI [30] dataset.

Table 4: Comparison of completion time of several algorithms (unit: seconds).

Algorithms	Data preprocessing time	Sampling time	Drawing time	Total time consumption
ICVC-GPU	15	25	26	66
MVRC2	14	22	19	56
SigIg+FMI	13	24	32	69
SHSR-UV	15	24	33	72
DARay	17	26	23	66
Ray-casting DAS	22	19	15	57

(2) Contrast reflects the influence of image on visual effect [32]. Contrast is proportional to the clarity of the image. The formula is as follows:

$$\mathrm{con} = L\sqrt{\frac{1}{MN}\sum_{x=0}^{L-1}\sum_{y=0}^{L-1}\left[I(x,y) - \frac{1}{MN}\sum_{x=0}^{L-1}\sum_{y=0}^{L-1} I(x,y)\right]^2}. \quad (21)$$

Among them, L denotes the average gray level of the pixels, M and N denote the width and height of the image, and $I(x,y)$ denotes the gray level of the pixels (x,y).

(3) Sharpness represents the clarity of the image, and the calculation formula is as follows

$$\mathrm{Def} = \sum_{x=0}^{M-2}\sum_{y=0}^{N-2} G(x,y). \quad (22)$$

(4) Firstly, the mean square error (MSE) [37] is defined. The mean square error is used to calculate the mean square value of the pixel difference between the

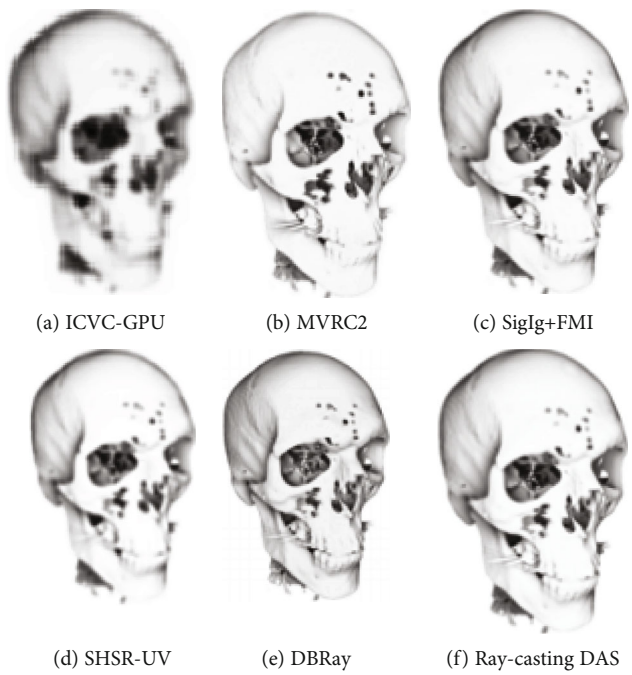


Figure 5: Comparison of the visual effects of the different algorithms in image rendering.

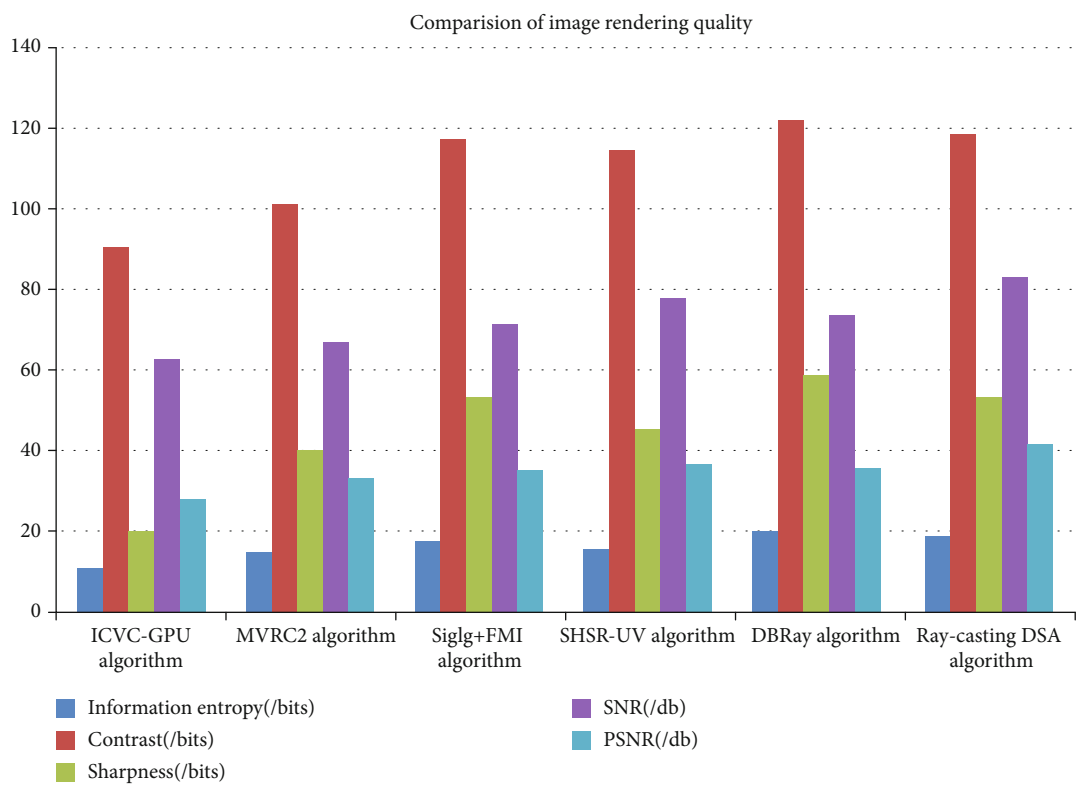


Figure 6: Comparison of image quality generated by several algorithms.

original image and the distorted image. The formula is as follows:

$$\mathrm{MSE} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}(R(m,n) - I(m,n))^2}{M \times N}. \quad (23)$$

Peak signal-to-noise ratio (PSNR) [32] is the ratio of maximum signal quantity to noise intensity. Since digital images are represented by discrete numbers, generally $L = 255$.

$$\mathrm{PSNR} = 10 \ln \frac{l^2}{\mathrm{MSE}} = 10 \ln \frac{255^2}{\mathrm{MSE}}. \quad (24)$$

(5) Signal to noise ratio (SNR)'s definition is as follows [31, 33]

$$\mathrm{SNR} = 10 \lg \left(\frac{\sum_{m=1}^{M} \sum_{n=1}^{N} R(m,n)^2}{\sum_{m=1}^{M} \sum_{n=1}^{N} \left(R(m,n) - I(m,n)\right)^2} \right). \quad (25)$$

The higher the SNR is, the less noise the image owns.

Through the experimental results in Figure 3, we can find that ray-casting DAS is slightly worse than algorithms E in image rendering quality, because the dynamic adaptive sampling used in this paper will inevitably skip some small voxels containing useful information due to the setting of the threshold value and other problems, leading to slightly poor effective information of the image, but it is close to or much better than A, B, C, and D in image rendering quality. Because the distance function is introduced into the process of rendering operator composition and the most appropriate distance function is selected, the image drawn by this algorithm has a high signal-to-noise ratio and contains less noise and other interference factors, compared with other algorithms. This is a significant progress. Simultaneously, the sampling and drawing time of ray-casting DAS, as shown in Table 3, is much better than all the other five algorithms, which indicates that our method is effective and saving time consumption.

5. Conclusion

Based on the low efficiency of the sampling point selection and slow rendering speed of the existing ray-casting algorithm, a new method based on dynamic sampling and improved rendering operator is proposed in this paper. The experimental results show that this method can effectively shorten the time of sampling and rendering and improve the efficiency of the algorithm on the premise of ensuring the high quality of graphics rendering.

At the same time, this paper introduces the concept of distance function in the process of image rendering, which makes the image contrast high and contains less noise. However, there are other shortcomings in this algorithm, such as the need for gradient calculation, and the gradient calculation is placed in the preprocessing stage, which aggravates the calculation amount in the preprocessing stage. It results in more time in the pretreatment stage. The efficiency of this algorithm can be further improved by a more efficient gradient calculation method.

Data Availability

The data used to support the findings of this study are available from the following websites: https://challenge2018.isic-archive.com/, https://grand-challenge.org/All_Challenges/, http://adni.loni.usc.edu/data-samples/accessdata/, and http://academictorrents.com/details/a9e2741587d42ef6139aa474a95858a17952b3a5.

Disclosure

The earlier version of this paper has been presented as a conference abstract in "ISAIR 2020: THE 5TH INTERNATIONAL SYMPOSIUM ON ARTIFICIAL INTELLIGENCE AND ROBOTICS 2020" according to the following link: https://easychair.org/smart-program/ISAIR2020/2020-08-09.html#talk:157507.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Campagnolo for publishing the dataset and Francisco Sans for making the code available. We are grateful for the comments from the anonymous reviewers. This research was supported in part by National Natural Science Foundation of China (Grant Nos. 61562013, 61772149, 61866009, 62002082), Guangxi Key Research and Development Project (Grant No. AB19110038), Guangxi Natural Science Foundation (Grant Nos. 2019GXNSFAA245014, 2017GXNFDA198025, 2018GXNSFAA294132, 2020GXNSFBA238014), and Guangxi University Young and Middle-aged Teachers' Research Ability Improvement Project (Grant Nos. 2019KY0238, 2020KY05034).

References

[1] A. Beristain, J. Congote, and O. E. Ruiz, "Volume visual attention Maps (VVAM) in ray-casting rendering," in *Medicine Meets Virtual Reality (MMVR)*, pp. 53–57, Newport Beach, California, USA, April 2012.

[2] Z. Lesar, "Real-time ray casting of volumetric data," in *IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON)*, pp. 45–62, Salamanca, Spain, March 2015.

[3] H. Ray, H. Pfister, D. Silver, and T. A. Cook, "Ray casting architectures for volume visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 5, no. 3, pp. 210–223, 1999.

[4] J. R. Mitchell, P. Dickof, and A. G. Law, "A comparison of line integral algorithms," *Computers in Physics*, vol. 4, no. 2, pp. 166–172, 1990.

[5] M. Levoy, "Volume rendering: display of surface from volume data," *IEEE Computer Graphics & Applications*, vol. 8, no. 3, pp. 29–37, 1998.

[6] Z. Jian, "The study of volume rendering techniques for medical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 7, pp. 463–473, 2005.

[7] Z. Liu, H. Seo, A. Castiglione, K.-K. R. Choo, and H. Kim, "Memory-efficient implementation of elliptic curve cryptography for the Internet-of-Things," *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 3, pp. 521–529, 2019.

[8] K.-K. R. Choo, C. Esposito, and A. Castiglione, "Evidence and forensics in the cloud: challenges and future research directions," *IEEE Cloud Computing*, vol. 4, no. 3, pp. 14–19, 2017.

[9] I. Demir and R. Westermann, "Vector-to-closest-point Octree for surface ray-casting," in *Vision, Modeling and Visualization(VMV)*, pp. 65–72, Aachen, Germany, October 2015.

[10] I. Herrera, C. Buchart, I. Aguinaga, and D. Borro, "Study of a ray casting technique for the visualization of deformable volumes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 11, pp. 1555–1565, 2014.

[11] A. Averbuch, G. Lifschitz, and Y. Shkolnisky, "Accelerating X-ray data collection using pyramid beam ray casting geometries," *IEEE transactions on image processing*, vol. 20, no. 2, pp. 523–533, 2011.

[12] B. Lee, J. Yun, J. Seo, B. Shim, Y.-G. Shin, and B. Kim, "Fast high-quality volume ray-casting with virtual samplings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1525–1532, 2010.

[13] Z. Wu, N. Wang, J. Shao, and G. Deng, "GPU ray casting method for visualizing 3D pipelines in a virtual globe," *International Journal of Digital Earth*, vol. 12, no. 4, pp. 428–441, 2019.

[14] J. Luo, G. Hu, and G. Ni, "Dual-space ray casting for height field rendering," *Journal of Visualization and Computer Animation*, vol. 25, no. 1, pp. 45–56, 2014.

[15] D. Tost, S. Grau, M. Ferre, and A. Puig, "Ray-casting time-varying volume data sets with frame-to-frame coherence," in *Visualization and Data Analysis 2006*, vol. 6060, pp. 505–522, San Jose, CA, USA, January 2006.

[16] F. Sans and R. Carmona, "A comparison between GPU-based volume ray casting implementations: fragment shader, compute shader, OpenCL, and CUDA," *CLEI Electronic Journal*, vol. 20, no. 2, pp. 643–668, 2017.

[17] R. Binyahib, T. Peterka, M. Larsen, K.-L. Ma, and H. Childs, "A scalable hybrid scheme for ray-casting of unstructured volume data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 7, pp. 2349–2361, 2019.

[18] C. Schulz and A. Zell, "Sub-pixel resolution techniques for ray casting in low-resolution occupancy grid maps," in *2019 European Conference on Mobile Robots (ECMR)*, pp. 1–6, Prague, Czech Republic, Czech Republic, August 2019.

[19] L. Tao, T. Bui, and H. Hasegawa, "Global ray-casting range image registration," *IPSJ transactions on computer vision and applications*, vol. 9, no. 1, pp. 122–138, 2017.

[20] S. Lim and B.-S. Shin, "A half-skewed Octree for volume ray casting," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 7, pp. 1085–1091, 2007.

[21] F. Sans and R. Carmona, "Volume ray casting using different GPU based parallel APIs," in *2016 XLII Latin American Computing Conference (CLEI)*, pp. 1–11, Valparaíso, Chile, October 2016.

[22] L. Xiao, C. Li, Z. Wu, and T. Wang, "An enhancement method for X-ray image via fuzzy noise removal and homomorphic filtering," *Neurocomputing*, vol. 195, no. 2, pp. 56–64, 2016.

[23] M. Mouzai, C. Tarabet, and A. Mustapha, "Low-contrast X-ray enhancement using a fuzzy gamma reasoning model," *Medical & Biological Engineering & Computing*, vol. 58, no. 6, pp. 1177–1197, 2020.

[24] J. P. Wang, F. Yang, and Y. Cao, "Cache-aware sampling strategies for texture-based ray casting on GPU," in *2014 IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 19–26, Paris, France, November 2014.

[25] https://challenge2018.isic-archive.com/.

[26] F. Yuan, "An interactive concave volume clipping method based on GPU ray casting with Boolean operation," *Computing and Informatics*, vol. 31, no. 3, pp. 551–600, 2012.

[27] L. Q. Campagnolo and W. Celes, "Interactive directional ambient occlusion and shadow computations for volume ray casting," *Computers & Graphics*, vol. 84, pp. 66–76, 2019.

[28] Y. Zhang, P. Gao, and X. Li, "A novel parallel ray-casting algorithm," in *2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Chengdu, China, 2016.

[29] https://grand-challenge.org/All_Challenges/.

[30] http://academictorrents.com/details/a9e2741587d42ef6139aa474a95858a17952b3a5.

[31] L. Lin, S. Chen, Y. Shao, and Z. Gu, "Plane-based sampling for ray casting algorithm in sequential medical images," *Computational and mathematical methods in medicine*, vol. 2013, Article ID 874517, 5 pages, 2013.

[32] M. Bozorgi and F. Lindseth, "GPU-based multi-volume ray casting within VTK for medical applications," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 3, pp. 293–300, 2015.

[33] A. Mastmeyer, T. Hecht, D. Fortmeier, and H. Handels, "Ray-casting based evaluation framework for haptic force feedback during percutaneous transhepatic catheter drainage punctures," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 3, pp. 421–431, 2014.

[34] http://adni.loni.usc.edu/data-samples/access-data/.

[35] R. Mehaboobathunnisa, A. A. H. Thasneem, and M. M. Sathik, "Fuzzy mutual information-based intraslice grouped ray casting," *Journal of Intelligent Systems*, vol. 28, no. 1, pp. 77–86, 2019.

[36] A. G. Moore, J. G. Hatch, S. Kuehl, and R. P. McMahan, "VOTE: a ray-casting study of vote-oriented technique enhancements," *International Journal of Human-Computer Studies*, vol. 120, no. 12, pp. 36–48, 2018.

[37] M. Fröhlich, C. Bolinhas, and A. Depeursinge, "Holographic visualisation and interaction of fused CT, PET and MRI volumetric medical imaging data using dedicated remote GPGPU ray casting," in *POCUS 2018, BIVPCS 2018, CuRIOUS 2018*, pp. 102–110, Granada, Spain, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8842227, 14 pages
https://doi.org/10.1155/2020/8842227

Research Article

Terrain Classification Algorithm for Lunar Rover Using a Deep Ensemble Network with High-Resolution Features and Interdependencies between Channels

Lanfeng Zhou, Ziwei Liu, and Wenfeng Wang

Shanghai Institute of Technology, Shanghai, China

Correspondence should be addressed to Lanfeng Zhou; lfzhou@sit.edu.cn and Wenfeng Wang; wangwenfeng@sit.edu.cn

Received 25 June 2020; Revised 18 August 2020; Accepted 3 September 2020; Published 14 October 2020

Academic Editor: Yin Zhang

For terrain classification tasks, previous methods used a single scale or single model to extract the features of the image, used high-to-low resolution networks to extract the features of the image, and used a network with no relationship between channels. These methods would lead to the inadequacy of the extracted features. Therefore, classification accuracy would reduce. The samples in terrain classification tasks are different from in other image classification tasks. The differences between samples in terrain classification tasks are subtler than other image-level classification tasks. And the colours of each sample in the terrain classification are similar. So we need to maintain the high resolution of features and establish the interdependencies between the channels to highlight the image features. This kind of networks can improve classification accuracy. To overcome these challenges, this paper presents a terrain classification algorithm for Lunar Rover by using a deep ensemble network. We optimize the activation function and the structure of the convolutional neural network to make it better to extract fine features of the images and infer the terrain category of the image. In particular, several contributions are made in this paper: establishing interdependencies between channels to highlight features and maintaining a high-resolution representation throughout the process to ensure the extraction of fine features. Multimodel collaborative judgment can help make up for the shortcomings in the design of the single model structure, make the model form a competitive relationship, and improve the accuracy. The overall classification accuracy of this method reaches 91.57% on our dataset, and the accuracy is higher on some terrains.

1. Introduction

Terrain classification is important in the driving process of a lunar rover, especially in complex terrain environments. There are two main directions for terrain classification. The first one is to classify the terrain by the vibration at which the rover through the ground. The second is to classify the terrain by the vehicle camera. The second method is more widely used than the first, because, in most situations, people need to predict the terrain in front of them.

The model proposed in this paper is based on the second direction. The base approach is to extract information from terrain images, such as spectra, colour, texture, and scale-invariant feature transform (SIFT) features [1–5]. The terrain can be accurately identified by that.

Although considerable research has been conducted on terrain classification in recent years, the bulk of this research is focused on man-made environments [6] or use more traditional algorithms for classification. Howard and Seraji [7] of the Jet Propulsion Laboratory did a lot of research on terrain description and terrain traversability estimation. Visual characteristics and fuzzy rules are used to estimate the traversability of the terrain. Firstly, the surrounding terrain images are obtained through vehicle cameras. The roughness, slope, discontinuity, hardness, and other information of the surrounding terrain are obtained through image analysis. Then, the type of terrain is judged according to the established fuzzy rules. This paper gives us clear classification rules and a normative reference. Iagnemma et al. [8] of the Massachusetts Institute of Technology

proposed an online terrain parameter estimation method. Lauro Ojeda of the University of Michigan did some new research on terrain classification and some terrain descriptions [9]. They used a fully connected neural network with only one hidden layer to classify the terrain. They divided the samples into five categories. They are gravel, grass, sand, and pavement dirt. The accuracy of this model reached 78.4%. He et al. [10] proposed a hierarchical classification approach. The Conditional Random Field (CRF) and the Bayesian Network (BN) are employed to incorporate prior knowledge, to facilitate SAR image classification. However, from DeepLab v3 [11] to DeepLab v3+ [12], the CRF block is replaced by complex neural networks. It means that neural networks can replace traditional machine learning methods in some tasks, and even neural networks will perform better.

In addition, with the continuous development of artificial intelligence, more and more intelligent algorithms, such as CNN and unsupervised learning, are used for terrain classification. Zeltner [13] used a deep convolutional neural network to implement a vision-based terrain classification. Park et al. [14] proposed a new classification network framework based on LSTM units and ensemble learning. Lu et al. [15] detected deep-sea images by using YOLO. Bai et al. [16] proposed an improved terrain classification method based on three-dimensional vibration information for terrain classification.

It is worth noting that most previous works focus on the extraction of salient features. The samples have significant differences between each other in those tasks. However, performances of the most previous models will be degraded when the differences between images are subtle.

Combined with the above analysis, a new terrain classification method based on the combination of convolutional neural network and ensemble learning is proposed. The main contributions of this paper are establishing interdependencies between the channels to highlight the image features and maintaining a high-resolution representation throughout the process to ensure the extraction of fine features. We get the channel descriptor by using global average pooling along the channel direction on feature maps after each convolution. This descriptor has a global receptive field. We take this descriptor as input to a fully connected network. The output has the same number of channels as input. The weights of this fully connected network are used to represent the interdependencies among various channels. The output is weighted channel-by-channel to previous features by multiplication to achieve a feature map with different channel weights. To achieve high-resolution features, we connect high-to-low resolution subnetworks in parallel. The information is exchanged in parallel multiresolution subnets. Information can be exchanged directly between the same resolutions. Information is exchanged from high-resolution to low-resolution by downsampling. Information is exchanged from low-resolution to high-resolution by upsampling. We discard low-resolution features because the differences between the samples are subtle. Meanwhile, a new activation function Mish [17] is used in the optimization of the model. Mish is a new activation function that is proposed by Diganta Misra. It performs better than ReLU on many datasets. Besides, we use a shallow neural network to integrate the output of each model to get the final results. Multiple models can be more effective to prevent overfitting than a single model. In addition, our method is easy to implement in practical applications. The remainder of this paper is organized as follows.

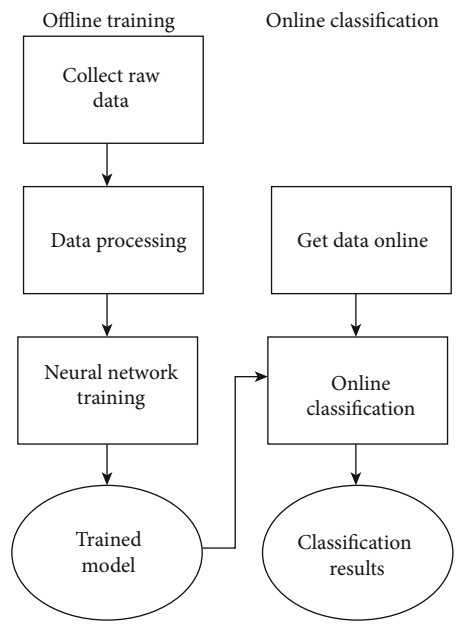


Figure 1: Schematic diagram of the algorithm flow framework.

In Materials and Methods, we review the previous methods, standards for terrain classification, and the theoretical basis of the convolutional neural network (CNN). We take these as the basis of our research. Meanwhile, we give our specific model. In Results and Discussion, we describe the experimental process of our model and the results of our experiments. In Conclusions, we summarize and analyse the experimental results.

2. Related Work

2.1. Terrain Classification. The main target of terrain classification is that we can quantify the ease-of-traversal of terrain by a mobile robot based on real-time measurements of terrain characteristics retrieved from vehicle cameras. Howard and Seraji [7] used a rule-based Fuzzy Traversability to classify the terrain. These characteristics include, but are not limited to slope, roughness, hardness, and discontinuity. The classification criteria of our experimental raw data are based on the above indicators.

2.2. Image Classification. In recent years, automatic classification techniques based on neural networks have been more and more. From 2012, on the competition of ImageNet [18], AlexNet [19] was proposed. On the ICLR2015, the VGG [20] was proposed. On the CVPR2018, Google proposed NasNet [21]; it was training on 500 GPUs. The accuracy of the top 5 and top 1 of the competition of ImageNet increased from 57.1% and 76.3% to 96.2% and 82.7%. It

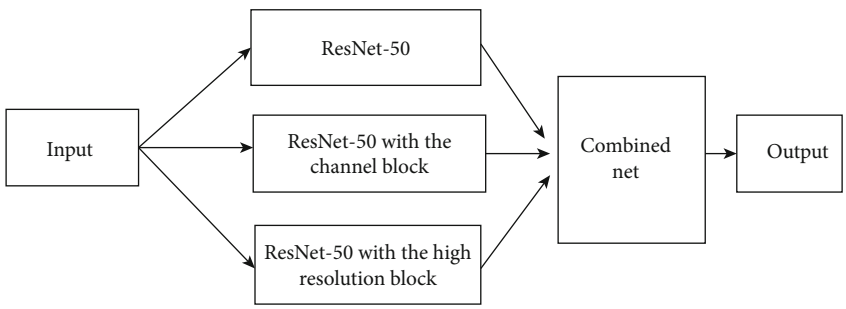


Figure 2: The structure of the ensemble network.

Table 1: A detailed description of the original ResNet-50 model.

conv1	7×7, 64, stride 2
	3×3, maxpool, stride2
conv2_x	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 256 \end{bmatrix} \times 4$
conv4_x	$\begin{bmatrix} 1 \times 12, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
conv5_x	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
fc1	Average pool, 1000-d fc, Softmax
fc2	4-d fc

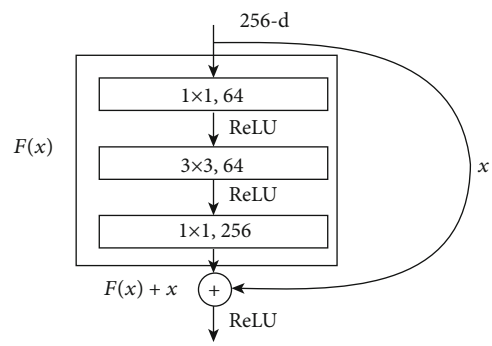


Figure 3: conv2_x structure.

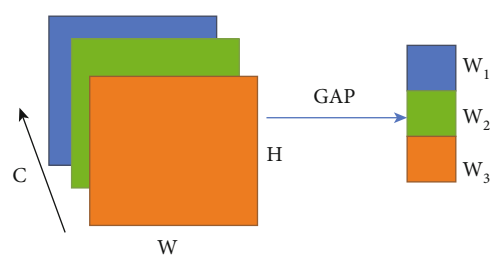


Figure 4: The process from the feature map to the initial descriptor of each channel.

can be seen that the accuracy of image classification based on convolutional neural networks is constantly improving.

2.3. Convolutional Neural Network. Neural networks were proposed by mimicking human brains. Convolutional neural networks are a special structure of neural networks, which are widely used in the field of computer vision. The image is divided into small regions in the same manner as the brain perceives the object. The features of each region are learned to classify the input image.

A simple neural network is a chained structure in which each layer is a function of the previous layer. [22, 23] The first layer:

$$H^{(1)} = g^{(1)}\left(W^{(1)T}x + b^{(1)}\right), \tag{1}$$

$H^{(1)}$ is the output of the first layer. $g^{(1)}$ is a nonlinear variation function. $W^{(1)}$ is the weight matrix, which the values we need to train. b is bias. x is an input layer. It is a vector.

The second layer:

$$H^{(2)} = g^{(2)}\left(W^{(2)T}H^{(1)} + b^{(2)}\right), \tag{2}$$

$H^{(1)}$ is the output of the first layer. So, we can express the output of n layer as:

$$H^{(n)} = g^{(n)}\left(W^{(n)T}H^{(n-1)} + b^{(n)}\right). \tag{3}$$

There is a theory that we can represent any arbitrary function by a neural network that has more than two layers. But according to experimental experience, training a deep network requires much fewer parameters than training a shallow network. The reason is that the low layers have already extracted the basic features, and the high layers only need to combine these basic features to get more complex

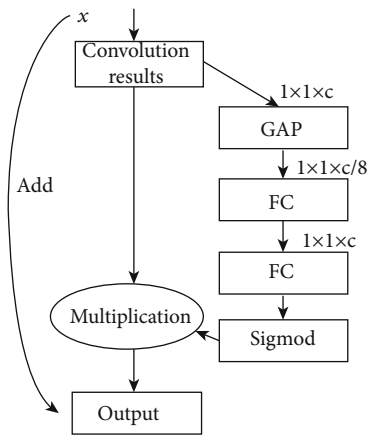


Figure 5: The structure of the fully connected network in the channel block.

Table 2: A detailed description of the ResNet-50 with channel block.

conv1	7×7, 64, stride 2	
	3×3 maxpool, stride 2	
conv2_x	conv, 1×1, 64 conv, 3×3, 64 conv, 1×1, 256 fc, [32, 256]	$\times 3$
conv3_x	conv, 1×1, 128 conv, 3×3, 128 conv, 1×1, 512 fc, [64, 512]	$\times 4$
conv4_x	conv, 1×1, 256 conv, 3×3, 256 conv, 1×1, 1024 fc, [128, 1024]	$\times 6$
conv5_x	conv, 1×1, 512 conv, 3×3, 512 conv, 1×1, 2048 fc, [256, 2048]	$\times 3$
fc1	Average pool, 1000-d fc, Softmax	
fc2	4-d fc	

features. It is similar to modularization in industrial production [24, 25].

3. Materials and Methods

3.1. Lunar Terrain Classification

3.1.1. Method Overview. The entire classification process is divided into two phases. The first part is the offline training.

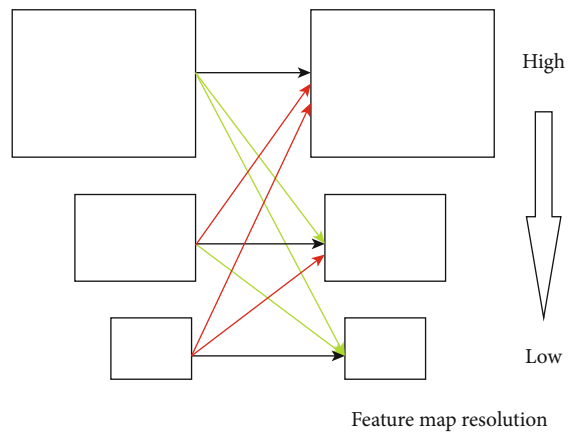


Figure 6: The process of exchanging information in three kinds of resolution in each channel.

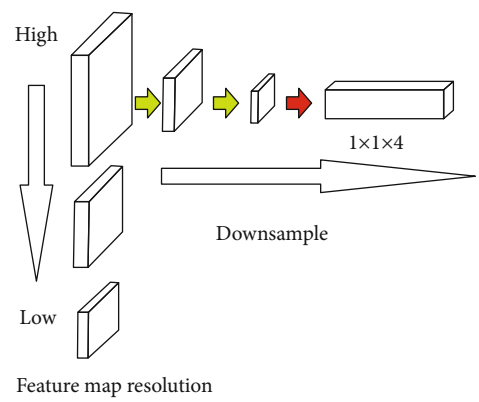


Figure 7: The process of constructing the classifier in three kinds of resolution.

The second part is the online classification. As shown in Figure 1, in the offline training part, we use more than 3,000 photos taken by Chang'e-3 as raw data. The data processing includes cleaning, labelling, dividing, and enhancing data. We train the model with our data. After a period of training, we get a trained model. In the online classification part, we get data online as inputs. Next, we use that trained model to classify the terrain. The trained model's outputs are the classification results. This is the whole process of the terrain classification algorithm.

3.2. Model. We use ResNet-50 [26] as a backbone and two functional blocks. One of the functional blocks is to establish the interdependencies between the channels. We call it the channel block. Another one is to maintain a high-resolution representation. We call it the high-resolution block. Our model includes 4 networks, original ResNet-50, ResNet-50 with the channel block, ResNet-50 with the high-resolution block, and a combined network. We divide samples into three groups and use three different models to train. Combined with three model results, the final results can be obtained. Ensemble learning is a technique that can alleviate overfitting problems. Figure 2 shows the structure of our ensemble network:

Table 3: A detailed description of the ResNet-50 with high-resolution block.

Resolution	Stage 1	Stage 2	Stage 3	Stage 4	Head
1	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 4 \times 1$	$\begin{bmatrix} 3\times3, 48 \\ 3\times3, 48 \end{bmatrix} \times 4 \times 1$	$\begin{bmatrix} 3\times3, 48 \\ 3\times3, 48 \end{bmatrix} \times 4 \times 4$	$\begin{bmatrix} 3\times3, 48 \\ 3\times3, 48 \end{bmatrix} \times 4 \times 3$	Two 3×3 convolution for downsample
1/2		$\begin{bmatrix} 3\times3, 96 \\ 3\times3, 96 \end{bmatrix} \times 4 \times 1$	$\begin{bmatrix} 3\times3, 96 \\ 3\times3, 96 \end{bmatrix} \times 4 \times 4$	$\begin{bmatrix} 3\times3, 96 \\ 3\times3, 96 \end{bmatrix} \times 4 \times 3$	
1/4			$\begin{bmatrix} 3\times3, 192 \\ 3\times3, 192 \end{bmatrix} \times 4 \times 4$	$\begin{bmatrix} 3\times3, 192 \\ 3\times3, 192 \end{bmatrix} \times 4 \times 3$	
1/8				$\begin{bmatrix} 3\times3, 384 \\ 3\times3, 384 \end{bmatrix} \times 4 \times 3$	

The following Table 1 is a detailed description of the original ResNet-50 [26] model.

The shapes and operations with specific parameter settings of a residual building block are listed inside the brackets and the number of stacked blocks in a stage is presented outside. The fc1 and fc2 mean two fully connected layers [27]. The ReLU activation function is not shown for brevity.

The convolution formula is as follows:

$$p_{i,j} = f\left(\sum_{d=0}^{D-1}\sum_{m=0}^{F-1}\sum_{n=0}^{F-1} w_{d,m,n} x_{d,i+m,j+n} + w_b\right) \tag{4}$$

$p_{i,j}$ is the pixel value of row i and column j, D is the depth of the convolution kernel, F is the size of the convolution kernel, $w_{d,m,n}$ is the weight of the convolution kernel in row m, and column $n\,c$; w_b is the bias.

ReLU activation function is as follows:

$$f(x) = \max(0, x) \tag{5}$$

is the output of the previous layer.

The linear calculation is as follows:

$$f = Wx + b \tag{6}$$

W is the weight of the model obtained through training, b is the bias term, and x is the output of the previous layer.

The Softmax function is as follows:

$$\text{softmax}(x_1, x_2, \cdots x_n) = \frac{e^{x_i}}{\sum_{i=1}^{n} e^{x_i}} \tag{7}$$

n is the number of types, and the Softmax value is the probability of each type. Obviously, the sum of all Softmax values is 1. A Softmax function is used to generate a label distribution containing 4 categories.

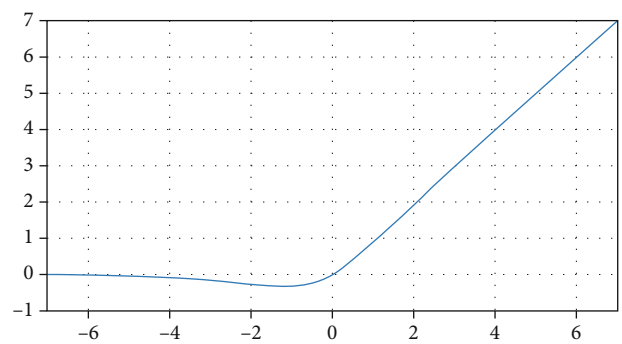


Figure 8: Mish function.

The cross-entropy:

$$\text{loss} = \sum_{i=1}^{n} p(x_i) \log(q(x_i)) \tag{8}$$

$p(x_i)$ is the true probability of x_i.$q(x_i)$ is the probability that is calculated by the model. The cross-entropy measures the distance between two distributions. The more similar the actual distribution to the predicted distribution, the smaller the value of the cross-entropy. Finally, we use Adam optimized gradient descent to solve this model. When the parameters converge, we can get the preliminary model.

3.3. Residual Block. To deal with the degradation problem, the residual block is proposed. The difference between ordinary neural networks is that the residual network has cross-layer connections, also called shortcut connections. A residual module is constructed by them. In a residual block, cross-layer connections generally span only two or three layers but do not exclude crossing more layers. The experimental results of the situation of crossing only one layer are not good. Figure 3 shows the conv2_x structure:

The residual block formula is as follows:

$$y = f(x, \{w_i\}) + x, \tag{9}$$

$f(x, \{w_i\})$is called residual mapping.

Table 4: A detailed description of the Combine Net.

Input	Layer 1	Layer 2	Result
ResNet-50			
ResNet-50 with channel block	Fully connected 512 neurons	Fully connected 4 neurons	Current terrain prediction results
ResNet-50 with high-resolution block			

The mapping between the l and l_1 layers is as follows:

$$a^{(l)} = f\left(a^{(l-1)}\right) + a^{(l-1)} = f\left(a^{(l-1)}\right) + f\left(a^{(l-2)}\right) + a^{(l-2)} = a^{(l_1)} + \sum_{i=l_1}^{l-1} f\left(a^{(i)}\right), \tag{10}$$

l, l_1 is any layer and $l > l_1$.

With the number of network layers deepening, the parameters of lower layers cannot be effectively updated in traditional networks. But in the residual block, we can solve it:

$$\frac{\partial \text{loss}}{\partial a^{(l_1)}} = \frac{\partial \text{loss}}{\partial a^{(l)}} + \frac{\partial \text{loss}}{\partial a^{(l)}} \frac{\partial}{\partial a^{(l_1)}} \sum_{i=l_1}^{l-1} f\left(a^{(i)}\right), \tag{11}$$

The gradient of the loss to a lower layer output is decomposed into two terms; the previous term $\partial \text{loss}/\partial a^{(l)}$ shows that error signals can propagate directly to lower layers without any intermediate weight matrix transformation. So the parameters of lower layers can be effectively updated. Residual connections make information flow more smoothly.

3.4. Channel Block. To establish the interdependencies between the channels, we use the channel block. Inspired by the 2017 ImageNet Challenge champion model [28], we find that the interdependencies between channels are useful to highlight features, especially for the classification model of samples with similar colours. We use the global average pooling(GAP) to get the descriptor of each channel. Figure 4 shows the process from the feature map to the initial descriptor of each channel. The number of channels is 3.

The global average pooling is as follows:

$$w_k = \sum_{i=0}^{W} \sum_{j=0}^{H} p_{i,j} / (W \times H), \tag{12}$$

w_k is the initial descriptor of the$k$ channel of the feature map. $p_{i,j}$ is the pixel value of row i and column j. W is the width of the feature map. H is the height of the feature map.

This descriptor has a global receptive field. Pixel context information is fully utilized. Getting this information is an important part of the picture-level classification.

We take the initial descriptors as the input of a fully connected network. The weights can represent the interdependencies between various channels by learning. The output is weighted channel-by-channel to previous features by multiplication to achieve a feature map with different channel weights. Figure 5 shows the structure of the fully connected network in the channel block.

Table 5: Rule base for Fuzzy Traversability Index.

Roughness	Discontinuity	Hardness	Type
Smooth	Small	Soft	Soft gravel
Smooth	Small	Hard	Compacted soil
Rough	Small	Hard	Compacted soil
Rough	Small	Soft	Soft gravel
	Large		Concave land
Rocky			Rocky terrain

FC is a fully connected layer. Sigmod is as the final activation function:

$$S(x) = \frac{1}{1 + e^{-x}}. \tag{13}$$

The output is a feature map with different channel weights.

Table 2 is a detailed description of the ResNet-50 with channel block.

The shapes and operations with specific parameter settings of a residual building block are listed inside the brackets and the number of stacked blocks in a stage is presented outside. The inner brackets following by fc indicates the output dimension of the two fully connected layers in a channel block.

The fc1 and fc2 mean two fully connected layers.

3.5. High-Resolution Block. To get subtle features, we use the high-resolution block. Inspired by the research of Human Pose Estimation in CVPR2019 [29], we find that extracting high resolution can improve the accuracy of the classification model, especially for the classification model of samples with only subtle differences. We use convolution to downsampling when information is exchanged from high resolution to low resolution. On the other hand, we use transposed convolution [30] to upsampling when information is exchanged from low resolution to high resolution. Figure 6 shows the process of exchanging information in three kinds of resolution in each channel.

The black arrow means identity mapping. The red arrow means transposed convolution. The green arrow means ordinary convolution. We see the original convolution as a

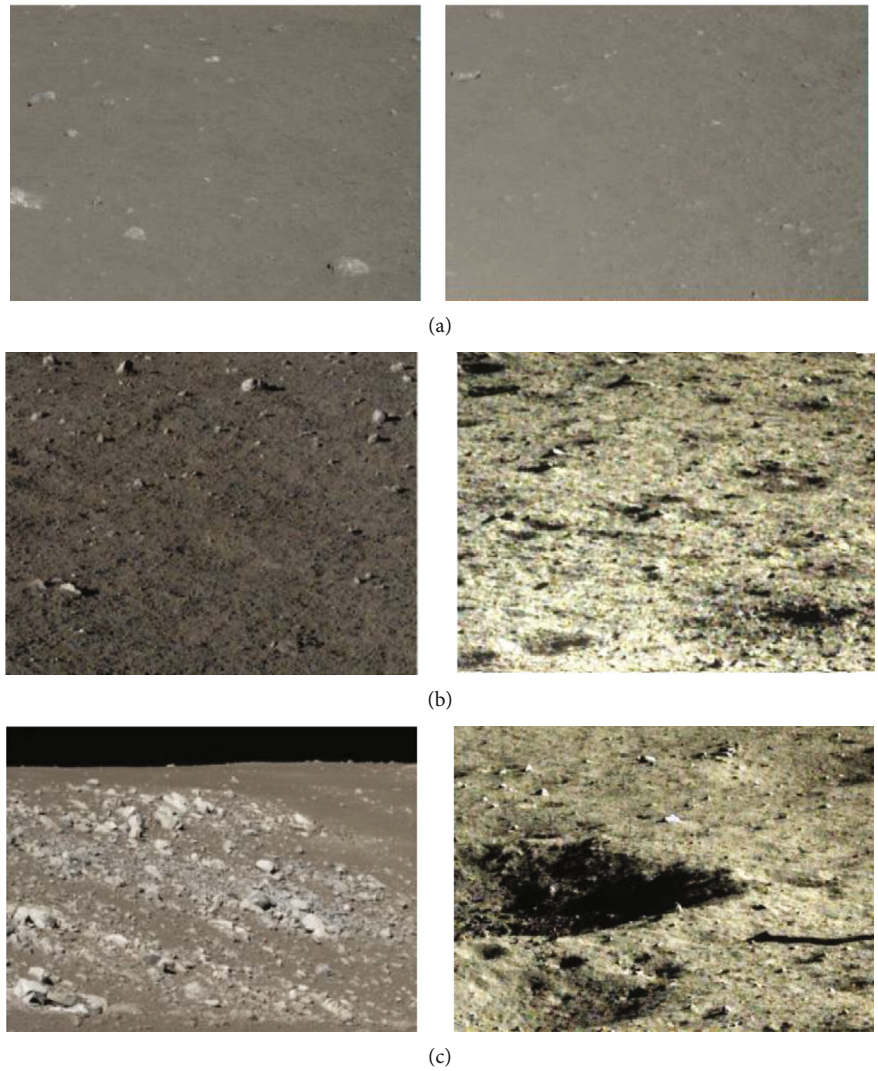


Figure 9: Raw data from different terrain. (a) Soft gravel. (b) Compacted soil. (c) Rocky terrain (left) and concave land(right).

Table 6: The distribution of data.

Class	Soft gravel	Compacted soil	Rocky terrain	Concave land
Number	1043	2765	625	368

matrix operation. If the size of the feature map is 4×4, the size of the convolution kernel is 3×3, the stride is 1, and then the size of the output of the convolution operation is 2×2. We can unroll the feature map, the output, and the convolution kernel into vectors from left to right, top to bottom. The convolution can be represented as follows:

$$W = \begin{pmatrix} w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{0,0} & w_{0,1} & w_{0,2} & 0 & w_{1,0} & w_{1,1} & w_{1,2} & 0 & w_{2,0} & w_{2,1} & w_{2,2} \end{pmatrix} \quad (14)$$

The feature map can be represented as follows:

$$x^T = \left(p_{0,0}p_{0,1}p_{0,2}p_{0,3}p_{1,0}p_{1,1}p_{1,2}p_{1,3}p_{2,0}p_{2,1}p_{2,2}p_{2,3}p_{3,0}p_{3,1}p_{3,2}p_{3,3}\right). \tag{15}$$

The output can be represented as follows:

$$S^T = \left(y_{0,0}y_{0,1}y_{1,0}y_{1,1}\right). \tag{16}$$

The original convolution can be seen as a matrix operation as follows:

$$W \cdot x = S. \tag{17}$$

The transposed convolution is as follows:

$$x = W^T \cdot S. \tag{18}$$

W^T is the kernel of transposed convolution.

Since our samples of classification tasks are with only subtle differences, the low-resolution features are not important. When constructing the classifier, we only use the high-resolution features. Figure 7 shows the process of constructing the classifier in three kinds of resolution.

The green arrow means that we use a 2-stride 3×3 convolution outputting 256 channels to downsampling the high-resolution feature map. The red arrow means that we use a 1-stride 1×1 convolution outputting 4 channels and GAP to put the representations into the classifier. 1×1 convolution is very useful to make the number of channels consistent or make the number of the channels any value we want.

Table 3 is a detailed description of the ResNet-50 with high-resolution block.

3.6. Combine Net. The following is a detailed description of the Combine Net:

The first layer of Combine Net is a simple fully connected layer with 500 neurons, making a simple nonlinear transformation of the results from the previous models.

The second layer is also a simple fully connected layer with only four neurons, making a simple nonlinear transformation of the results from the previous layer.

The activation function of the first layer is a Mish layer.

The activation function of the second layer is a Softmax which gives the final probability of each type. Mish is a novel smooth and nonmonotonic neural activation function which can be defined as:

$$f(x) = x \cdot \tan h(\ln(1 + e^x)) \tag{19}$$

The graph of Mish is shown in Figure 8.

Table 4 is a detailed description of the Combine Net.

3.7. Data Processing

3.7.1. Data Augmentation Transformations. We use more than 3,000 photos taken by Chang'e-3 as raw data. After cleaning and cutting, we get 4,801 valid samples. The size of the samples is $784 \times 576 \times 3$. According to the rule-based Fuzzy Traversability, 4801 images are labelled. The rule-based definition of the Traversability Index in terms of terrain roughness, discontinuity, and hardness is summarized in Table 5.

Table 7: Data augmentation transformations.

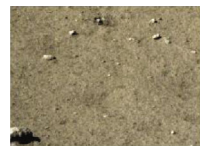
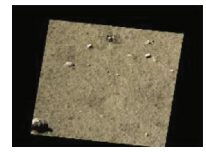
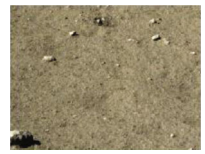
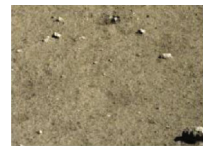
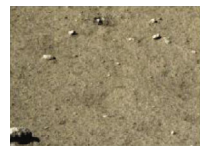
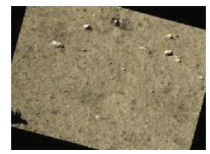
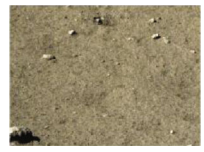
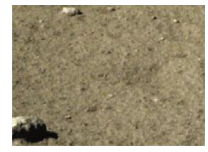
Transformation	Original	Image transform
Shear		
Flip right-left		
Rotate		
Scale		

Table 8: The number of images after the transformation.

Class	Soft gravel	Compacted soil	Rocky terrain	Concave land
Number	2086	2034	2187	2208

According to Table 5, the 4,801 sample images were divided into four categories: soft gravel topography, compacted soil topography, rocky terrain, and concave land. Among them, the compacted soil topography is the optimal travel choice, and the soft gravel topography has a large slip ratio. Rocky terrain and concave terrain should be avoided choosing to travel as much as possible. The following four sets of images in Figure 9 show the four types of samples.

The distribution of data is shown in Table 6:

After that, we perform a series of transformations on the image as shown in Table 7 to keep the distribution of the data balanced and alleviate overfitting problems.

We transform rocky terrain, soft gravel concave, and land to increase data. We discard some compacted soil to keep the distribution of the data balanced. The number of each terrain samples is shown in Table 8 after the transformation.

3.7.2. Read the Large-Scale Datasets. We build a convolutional network model using the TensorFlow-GPU-1.12.0. The GPU is Titan. The CPU is Xeon6130. In the training process, we divide the data into three equal parts and train the

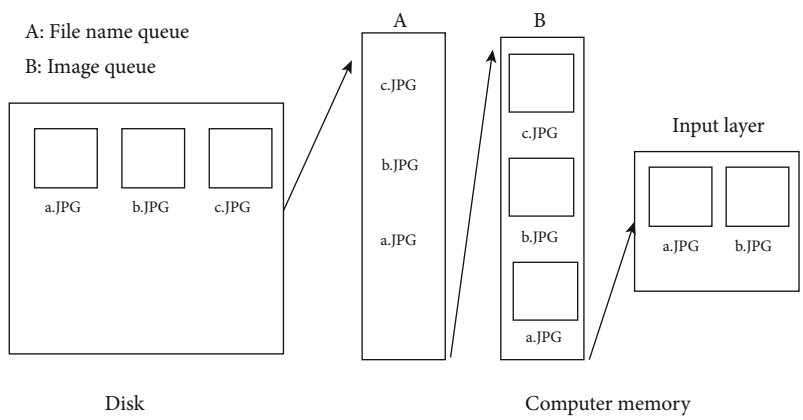


Figure 10: Picture reading process.

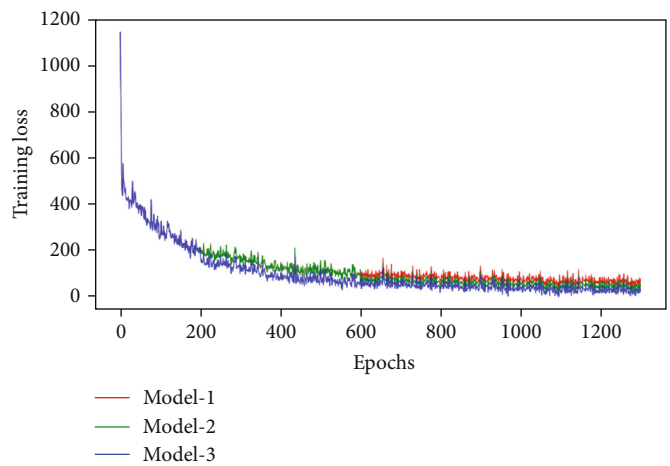


Figure 11: Loss values.

original ResNet-50 model, the ResNet-50 with the channel block model, and the ResNet-50 with the high-resolution model. The data are divided into the training set, the test set, and the verification set according to the ratio of 7 : 2 : 1. Every 16 pictures were used as a training batch. According to this, each batch is the input of the convolutional neural network for training, considering that I/O operations take a longer time than the encoding process and matrix operations. But we cannot put all the training data into computer memory because we do not have enough memory to store them. In order not to waste GPU resources, we use the multi-threaded operation to train the model, as Figure 10 shows,

Specifically, a thread is used to continuously read the image name and path from the disk into the file name queue. A thread is used to continuously read images from the disk according to the name in the file name queue into the image queue. The input layer of the model can get images continuously from the image queue.

4. Results and Discussion

4.1. Experimental Process and Results. First, we read the data from the image queue.

Then, we divided the data into three equal parts. We put every batch of images into the corresponding network and train 1300 epochs. Model-1 is the original ResNet-50 model. Model-2 is the ResNet-50 with the channel block model. Model-3 is the ResNet-50 with the high-resolution model.

The loss values of the verification set are shown in Figure 11.

The accuracy of the verification set is shown in Figure 12.

According to Figures 11 and 12, we found that model-3 gets the best results. It means high-resolution features are important in the terrain classification task which the samples in the task only have subtle differences. Model-2 gets better results than model-1. It means establishing the interdependencies between the channels can highlight the features.

Finally, we use the idea of ensemble learning to fuse three models and then train a simple neural network based on the results of the three models. We take its output as the final output. This process is similar to a voting process, but the weights of each vote are nonlinear. We call the ensemble network as model-ensemble.

The accuracy of the validation set is shown in Figure 13.

Compared to a single model, the ensemble model is more accurate than a single model. Table 9 shows the results.

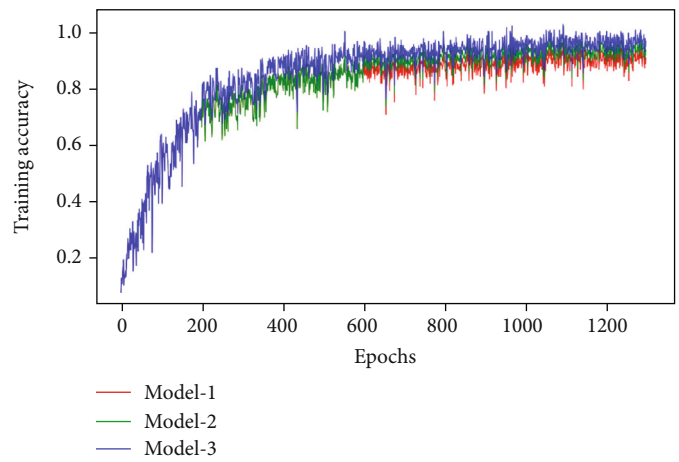


Figure 12: Accuracy values.

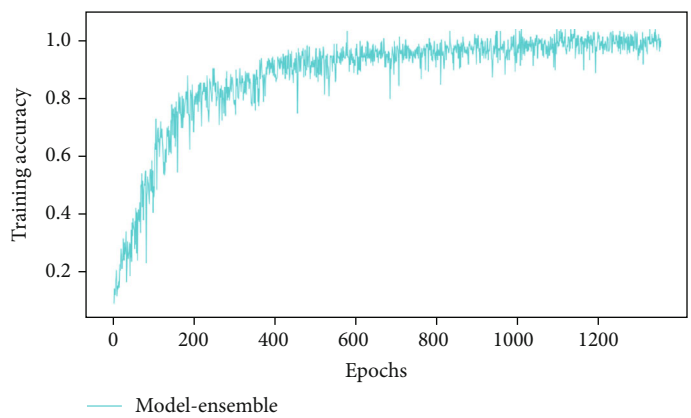


Figure 13: Accuracy of the ensemble model.

Table 9: Accuracy of the ensemble model.

Model	Train set	Test set	Val set
Model-1	92.61	91.07	89.75
Model-2	93.49	92.77	91.49
Model-3	94.11	93.16	91.54
Model-ensemble	94.31	93.24	91.57

Obviously, we see the result of model-ensemble is better than the result of model-3. The average accuracy is 91.57%. For the compacted soil terrain, the accuracy of our model is 95.37%. The accuracy of the soft gravel terrain is 93.95%, the accuracy of the rocky terrain is 89.13%, and the accuracy of the concave land terrain is 87.83%. The reason for the poor accuracy of the concave land terrain and the rocky terrain is that there are only 368 concave land images and 625 rocky terrain images in our sample. Even though we have enhanced the data, the results are still not good. Conversely, there are 2,765 compacted soil images. The accuracy of the compacted soil terrain is high. This indicates that the model requires a large number of samples to learn. In this situation, the model can learn to extract features and classify the terrain correctly.

From the above results, the imbalance of the data samples will cause the model to have different degrees of accuracy for different types of terrain. ROC curves and AUC values in Figure 14 intuitively express the ability of the model to recognise different terrains.

ROC curve of class 0-3 represents the ability of the model to recognize soft gravel, compacted soil, rocky terrain, and concave land. The average ROC curve is the average of the other 4 ROC curves. The larger the AUC values, the better the model. So the model is good at recognizing compacted soil. The reason is that there are more original data than others.

Considering the randomness of a single experiment, there are 20 random experiments to verify the model. The data is randomly selected from the data set. The results are shown in Figure 15. From the final results, the classification accuracy of the four types of terrain is 93.95%, 95.37%, 89.13%, and 87.83%, respectively. The results show that the classification accuracy of compacted soil and soft gravel is better than that of rocky terrain and concave land. Combined with the recognition accuracy of the whole dataset, it can be seen that the low accuracy may be caused by the difference in the amount of sample data. In the whole dataset, the amount of rocky terrain and concave land is smaller than that

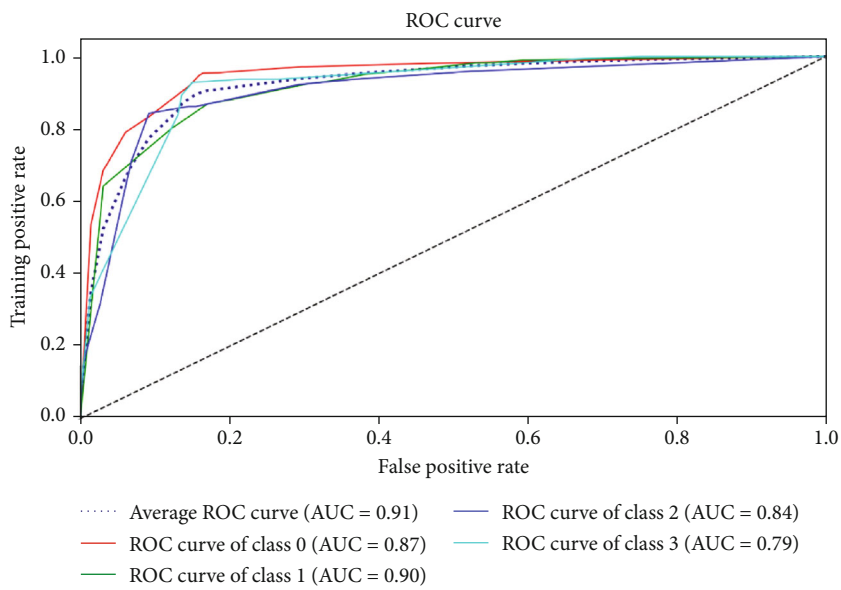


Figure 14: ROC curves and AUC values.

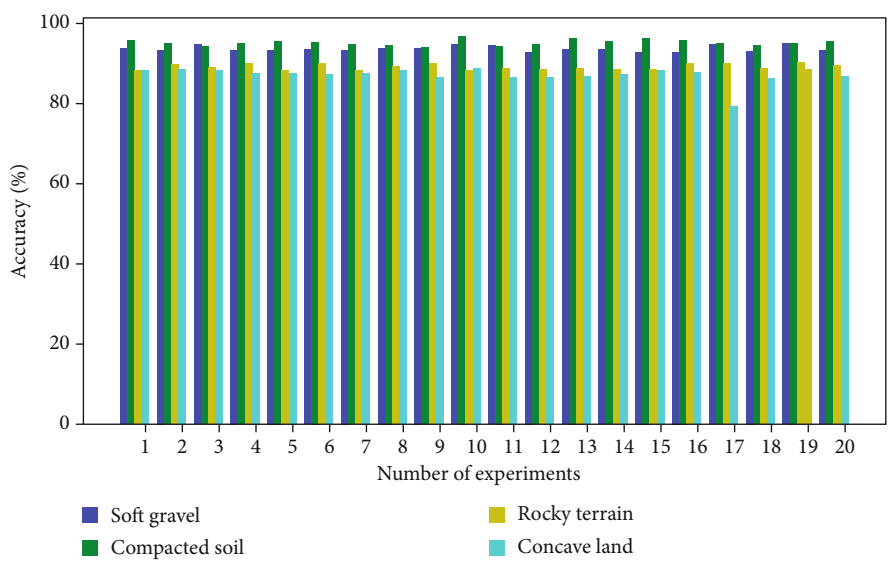


Figure 15: 20 experimental comparison results.

of soft gravel and compacted soil, which makes the learning accuracy inadequate.

Through the analysis of the confusion matrix of the experiment, 200 random data were classified into four types of terrain shown in Figure 16. It can be seen that the probability of rocky terrain and concave land being misclassified into wrong types of terrain is higher, which is also the reason for its low accuracy. We will be analyzing it in-depth in the follow-up study.

Since the geological composition of the moon is similar to that of the earth, we randomly searched for some scenes on the earth and used our model to classify. The results of the accuracy reached 90%. There are three samples of that random experiment in Table 10. The results mean that our model is valid to moon ground environment.

5. Conclusions

In this study, we proposed a method of the terrain classification algorithm for Lunar Rover by using a deep ensemble network with high-resolution features and interdependencies between channels. We use the original ResNet-50 model, the ResNet-50 with the channel block model, the ResNet-50 with the high-resolution model, and the Combine Net with a new activation function to solve terrain classification problems. To verify the algorithm, the experiment compares the performance of every single model on datasets and the performance of the deep ensemble network on datasets. The experimental results show that high-resolution features are important in the terrain classification task which the samples in the task only have subtle differences, establishing the

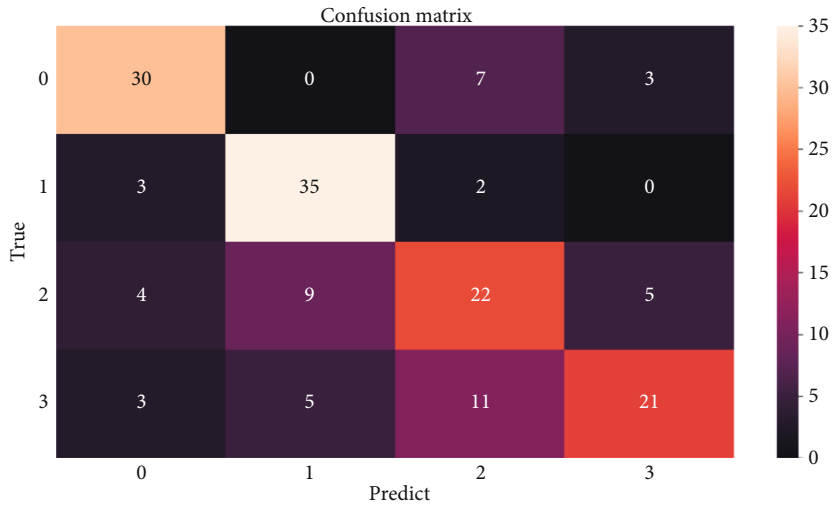


Figure 16: The results of classifying 200 random data in four types of terrain.

Table 10: The results of the random experiment.

Real scene	Classification result	True/false
	Soft gravel	True
	Compacted soil	True
	Rocky terrain	True

interdependencies between the channels can highlight the features, and multimodel collaborative judgment can help make up for the shortcomings in the design of the single model structure. Finally, the deep ensemble network reaches 91.57% average accuracy on our datasets.

Data Availability

You can get the raw data from here: http://moon.bao.ac.cn/

Conflicts of Interest

There is no conflict of interest regarding the publication of this paper.

Acknowledgments

The authors thank the research team for their support. This paper is funded by the National Natural Science Foundation (41671402).

References

[1] L. Semler and J. Furst, "Wavelet-based texture classification of tissues in computed tomography," in *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, pp. 265–270, Dublin, Ireland, 2005.

[2] G. Paschos, "Perceptually uniform color spaces for color texture analysis: an empirical evaluation," *IEEE Transactions on Image Processing*, vol. 10, no. 6, pp. 932–937, 2001.

[3] X. Liu and D. Wang, "Texture classification using spectral histograms," *IEEE Transactions on Image Processing*, vol. 12, no. 6, pp. 661–670, 2003.

[4] M. Pietikäinen, T. Mäenpää, and J. Viertola, *Color Texture Classification with Color Histograms and Local Binary Patterns*, IWTAS, New York, NY, USA, 2002.

[5] S. Zenker, E. E. Aksoy, and D. Goldschmidt, "Visual terrain classification for selecting energy efficient gaits of a hexapod robot," in *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pp. 577–584, Wollongong, NSW, Australia, July 2013.

[6] J. Kruger, A. Rogg, and R. Gonzalez, "Estimating wheel slip of a planetary exploration rover via unsupervised machine learning," in *2019 IEEE Aerospace Conference*, pp. 1–8, Big Sky, MT, USA, March 2019.

[7] A. Howard and H. Seraji, "Vision-based terrain characterization and traversability assessment," *Journal of Robotic Systems*, vol. 18, no. 10, pp. 577–587, 2001.

[8] K. Iagnemma, H. Shibly, and S. Dubowsky, "On-Line Terrain Parameter Estimation for Planetary Rovers," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, pp. 3142–3147, Washington, DC, USA, 2002.

[9] L. Ojeda, J. Borenstein, G. Witus, and R. Karlsen, "Terrain characterization and classification with a mobile robot," *Journal of Field Robotics*, vol. 23, no. 2, pp. 103–122, 2006.

[10] C. He, X. Liu, D. Feng, B. Shi, B. Luo, and M. Liao, "Hierarchical terrain classification based on multilayer Bayesian network and conditional random field," *Remote Sensing*, vol. 9, no. 1, p. 96, 2017.

[11] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, http://arxiv.org/abs/1706.05587.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, pp. 801–818, 2018.

[13] F. Zeltner, *Autonomous Terrain Classification through Unsupervised Learning*, University of Wurzburg, Master thesis, 2016.

[14] J. Park, K. Min, H. Kim, W. Lee, G. Cho, and K. Huh, "Road surface classification using a deep ensemble network with sensor feature selection," *Sensors*, vol. 18, no. 12, p. 4342, 2018.

[15] H. Lu, D. Wang, Y. Li et al., "CONet: a cognitive ocean network," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 90–96, 2019.

[16] C. Bai, J. Guo, and H. Zheng, "Three-dimensional vibration-based terrain classification for mobile robots," *IEEE Access*, vol. 7, pp. 63485–63492, 2019.

[17] D. Misra, "Mish: A Self Regularized Non-Monotonic Neural Activation Function," 2019, http://arxiv.org/abs/1908.08681.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, FL, USA, June 2009.

[19] A. Krizhevsky, I. Sutskever, and G. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, NIPS. Curran Associates Inc., 2012.

[20] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014, http://arxiv.org/abs/1409.1556.

[21] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018.

[22] S. M. Ahn, "Deep learning architectures and applications," *Journal of Intelligence and Information Systems*, vol. 22, no. 2, pp. 127–142, 2016.

[23] L. Ogiela and M. R. Ogiela, "Beginnings of cognitive science," in *Advances in Cognitive Information Systems. Cognitive Systems Monographs, vol 17*pp. 1–18, Springer, Berlin, Heidelberg.

[24] F. Seide, G. Li, and D. Yu, "Conversational speech transcript using context-dependent deep neural networks," in *INTERSPEECH 2011 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, August 2011.

[25] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., pp. 818–833, Springer, Cham, 2014.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[27] P. Simard, D. Steinkraus, and J. Platt, "Best practices for convolutional neural networks applied to visual document analysis," *ICDAR*, vol. 2, p. 958, 2003.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[29] S. Ke, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5693–5703, 2019.

[30] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016, http://arxiv.org/abs/1603.07285.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8813615, 13 pages
https://doi.org/10.1155/2020/8813615

Research Article

A Novel Search Ranking Method for MOOCs Using Unstructured Course Information

Weiqiang Yao,[1] **Haiquan Sun**,[1,2] **and Xiaoxuan Hu**[1,2]

[1]*School of Management, Hefei University of Technology, Hefei, Anhui 230009, China*
[2]*Key Laboratory of Process Optimization and Intelligent Decision Making, Ministry of Education, Hefei, Anhui 230009, China*

Correspondence should be addressed to Weiqiang Yao; wqyao@ustc.edu.cn

Received 25 June 2020; Revised 15 August 2020; Accepted 8 September 2020; Published 23 September 2020

Academic Editor: Yin Zhang

Massive open online courses (MOOCs) are a technical trend in the field of education. As the number of available MOOCs continues to grow dramatically, the difficulty for learners to find courses that satisfy their personalized learning goals has also increased. Unstructured texts, such as course descriptions and course skills, contain rich course information and are useful for MOOC platforms in constructing personalized services. This paper proposes a novel search ranking method for MOOCs that integrates unstructured course information. We propose a latent Dirichlet allocation-based model to cluster courses into groups based on course descriptions. Courses in the same cluster are considered to share similar educational contents. We then propose the CourseRank algorithm based on the information of course skills to recommend and rank courses when students search for or click on a specific course. Our experiments on the dataset from Coursera indicate that our method is able to cluster courses effectively and produce satisfactory ranking results for courses in MOOC platforms.

1. Introduction

Massive open online courses (MOOCs) have gained considerable global attention in the field of education. It offers a new way for organizations to share their knowledge and offer world-class education to the public [1]. A survey by Class Central shows more than 900 universities around the world launched more than 11.4 thousand MOOCs in various MOOC platforms in 2018 [2]. The number of students enrolled in MOOCs increased from 78 million in 2017 to more than 101 million in 2018.

With the increasing popularity of MOOCs, hosting as many courses as possible to satisfy various demands from students is a profitable business strategy for MOOC platforms. However, a common issue for MOOC platforms is that many courses in a platform have similar titles but different technical contents. Many courses with different titles may also have similar content because they cover the same knowledge points. Take http://Coursera.com/, for example; when we search the keyword "machine learning," more than 100 courses show up with titles containing the keyword "machine learning." These courses, such as "TensorFlow in Practice," in the list of search results also include related knowledge points although their titles do not have the keyword "machine learning." In such a case, if a student wants to learn certain knowledge or skills, the large number of similar courses makes it difficult for students to choose the right courses and achieve their personalized learning objectives. From the perspective of the platform, designing methods to assist students in finding MOOCs that can satisfy their learning objectives are necessary.

Many methods have been proposed to construct selection and ranking models for MOOCs. For example, Bousbahi and Chorfi [3] designed the case-based reasoning (CBR) approach and information retrieval technique to recommend MOOCs for learners. Elbadrawy and Karypis [4] investigated how student characteristics and course features affect course enrollment patterns. In these studies, the structured demographic characteristics, study records, and course features are the main information used to infer the learning

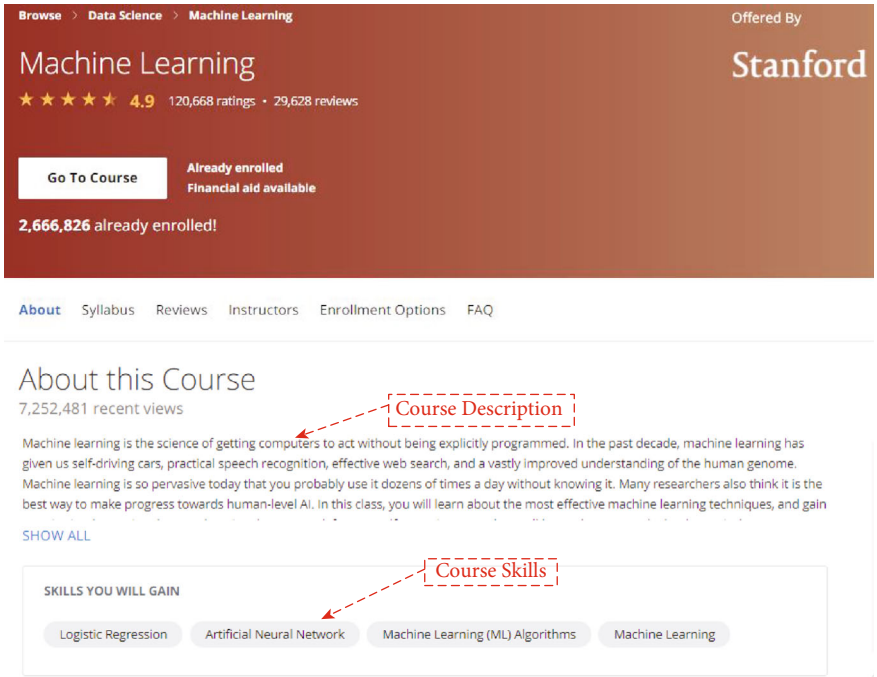


Figure 1: Unstructured information for a MOOC.

preferences of students. However, a large amount of unstructured data has not been explored fully to analyze student behaviors and provide personalized services.

In the MOOC platform, unstructured textual data, such as course descriptions, course skills, and student reviews usually imply useful course features. For example, http://Coursera.com/ (Figure 1) uses "About this Course" to introduce a specific course. Teachers can also present "skills you will gain" to indicate the contents and methods to be delivered in the course. The course description and skills help students know the teaching contents. Students can thus evaluate whether a course can meet their learning objectives [5]. For course search services, utilizing the textual information is useful for platforms to understand both the courses' teaching contents and students' learning objectives.

This paper designed a novel search ranking method for MOOCs with the unstructured course description released in MOOC platforms and the course skills given by teachers. When students click on a specific course or search some keywords in a MOOC platform search engine, we propose a model that can analyze the unstructured textual information and present students with a sorted list of courses. In the proposed method, the first stage is a latent Dirichlet allocation-(LDA-) based model to cluster courses into groups. The courses in the same clusters are considered to cover similar knowledge because they have comparable learning topics. All MOOC platforms offer many courses, and thus, each course cluster obtained in the first stage usually includes many courses. Hence, presenting all courses in the same cluster to students when they search for a keyword or click on a specific course is unreasonable and impractical. The second stage is a CourseRank algorithm to rank the courses in a cluster with the unstructured course skills. Courses with higher rankings are then selected and presented to the students. In general, the contributions of this study are three-fold:

(1) Technology-enhanced learning is a promoting trend in the field of education. Daniel suggested that working with big data and data science requires specialized skills lacking in many educational researchers [6]. This paper introduces machine learning technologies (i.e., LDA and PageRank algorithm) to the field of education research. The proposed models benefit research in the field of education by providing new technologies and tools to help researchers work with big data and data science. This study is valuable because it can help understand learners' cognition by analyzing the unstructured course information and can increase business efficiency for the MOOC platforms

(2) We employ the unstructured course description released in MOOC platforms and the course skills given by teachers for the course ranking algorithm. Although the unstructured course information contains rich knowledge on learning and teaching objectives, researches that have integrated the unstructured textual information are minimal, especially course skills information, into the course search ranking problem

(3) Instead of segmenting courses by clustering description words, the LDA-based model clusters the courses by extracting latent topics implied in the contents. This strategy can improve the results of

the course clustering and help platforms filter out unrelated courses to meet the individual preferences of students

The remainder of our research is organized as follows. Section 2 reviews the related work in literature. In Section 3, we propose the course clustering model and the course ranking model. In Section 4, we conduct experiments on the dataset from Coursera to test our proposed method. Section 5 concludes our research and provides the future directions.

2. Related Work

In this section, we review the previous works on MOOCs relevant to our study. We review the literature on student behaviors in the MOOC environment and the machine learning methods for MOOC ranking.

2.1. Student Behavior in the MOOC Environment. In the educational research, MOOC has drawn wide attention from scholars because it has been considered as one of the most effective online learning forms [7]. Bodily et al. regarded MOOC as one of the most important trends for instructional design and technology [8]. Zhu et al. reviewed MOOC research from 2014 to 2016 and classified current researches into several categories [9]. Costello et al. conducted a systematic review of research about the role of Twitter in the context of MOOCs from 2011 to 2017 [10]. Summarizing these literature reviews and current researches on MOOCs, student behavior is seen to be the most popular topic in literature, and current research generally used survey data to analyze student behaviors by descriptive statistics.

To study student behaviors in MOOCs environment, many scholars focused on student engagement in courses. For example, Aparicio et al. proposed a theoretical framework to identify the factors impacting MOOC use and satisfaction and empirically measure these factors in a real MOOC context [11]. Deng et al. developed and validated a MOOC engagement scale to measure learner engagement [12]. They found that behavioral engagement, emotional engagement, cognitive engagement, and social engagement are the four dimensions of student engagement in MOOCs. By taking into account factors such as expectancies, values, and social influence, Luik et al. studied factors that motivate the enrolment of learners in programming MOOCs [13]. Their study showed that interest in the course and personal suitability is the highest-rated motivational factors. Social influence and usefulness related to certification are the lowest-rated factors. Current literature investigated student engagement in MOOCs from the perspective of self-determination theory and the theory of relationship quality [14].

Aside from investigating student engagement, current literature also studied the learning behaviors of students after they enrolled in MOOC platforms. Cohen et al. characterized the active learners in forums and found that the completion status of learners significantly correlates to their activity in the forums [15]. Hood et al. examined how the current role and context of learners influence their ability to self-regulate their learning in the MOOC environment [16]. Significant differences were identified between learners with different characteristics. Guo and Reinecke studied the navigation behavior of students in the learning process [17]. Their results indicated that older students and those from countries with smaller student-teacher ratios are more comprehensive and nonlinear when navigating through the course.

The related works reviewed above indicate that most existing studies on MOOCs are empirical studies that use surveys or interview data [18]. New data sources and new methodologies are required to analyze learning behaviors in the MOOC environment. The literature review indicated that students with various characteristics often have different learning preferences and behaviors. Therefore, the MOOC platform needs a design operative strategy to predict student preference and provide suitable courses [19].

2.2. Machine Learning for MOOC Ranking. In the past several years, machine learning methods have been applied gradually to address issues in the field of MOOC research. Researchers employed methods such as random forest (RF), support vector machine (SVM), and LDA to understand student behaviors [20]. For example, Peng and Aggarwal transformed the MOOC dropout problem as a classification issue and designed several machine learning models based on SVM, gradient boosting decision trees, AdaBoost, and RF to solve the problem [21]. LDA is a popular text mining method for MOOCs. Ramesh et al. designed a seeded LDA model to understand MOOC discussion forums [22]. Atapattu and Falkner proposed an LDA-based framework to generate and label discussion topics automatically [23].

Course recommendation is an important research topic that emphasizes the employment of machine learning methods in the MOOC environment. Guo and Reinecke suggested that the function of course recommendation is necessary for MOOC platforms because it can help platforms provide proper courses to students and incentivize them to engage with the study process [17]. Hence, Bousbahi and Chorfi designed a MOOC recommendation method using CBR, which can effectively find the best learning resources for students [3]. Elbadrawy and Karypis proposed a domain-aware method to recommend courses based on the academic features of student and course groups [4]. Pang et al. proposed a multilayer bucketing recommendation method to recommend courses on MOOC platforms and designed a map-reduced technique to improve recommendation efficiency [24].

The above literature indicates that machine learning is one of the most popular methodologies in education research. However, although existing methods are useful, they usually rank courses by analyzing structured learning records or learner features. The unstructured data such as course descriptions and tags are yet to be explored. This paper employed LDA and PageRank to generate reasonable search results in the MOOC platforms. LDA and PageRank are machine learning methods widely used in various fields [25]. This study utilized the LDA algorithm to analyze course descriptions and cluster courses into groups, whereas the

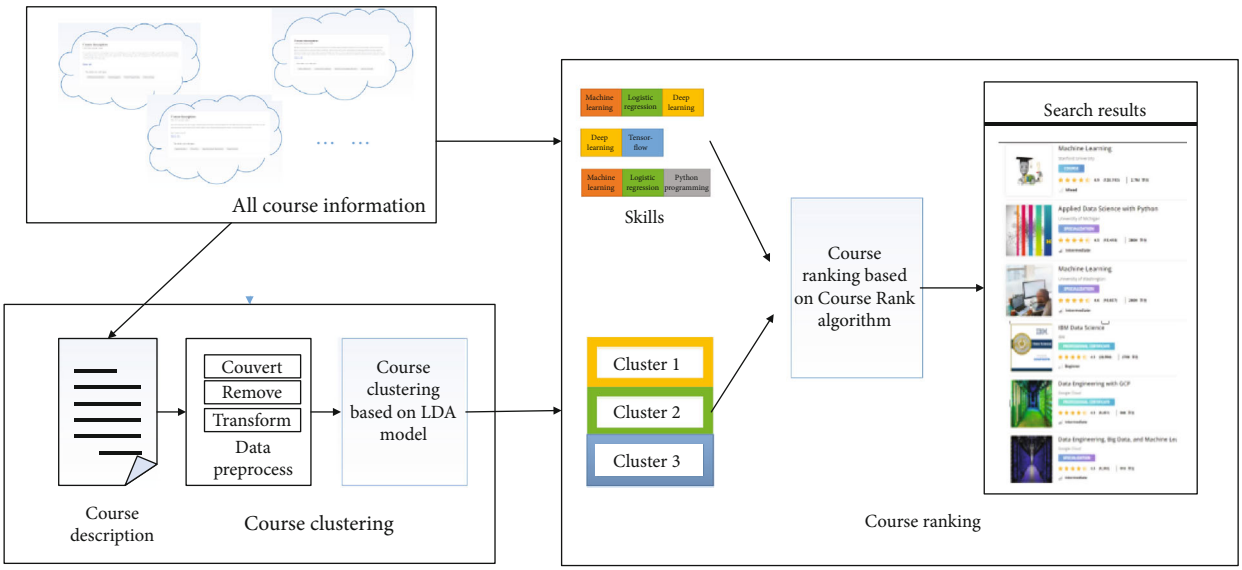


FIGURE 2: Framework of course ranking method.

PageRank algorithm was used to rank the courses in the same clusters. The proposed method is detailed next.

3. Search Ranking Method for MOOCs

In this section, we propose the search ranking method for courses based on the LDA and PageRank algorithm. Figure 2 provides the framework of our search ranking method. Figure 2 shows that based on the textual description information, we design a LDA-based model to cluster courses. For the courses in each cluster, a course ranking algorithm is proposed based on the skills which will gain through the courses. We provide the details of the proposed search ranking method for the course in the following sections.

3.1. Stage 1: LDA-Based Model for Course Description Clustering. We now provide the LDA-based model for course clustering. LDA uses an unsupervised Bayesian model to capture context-specific dimensions implied in the unstructured course description. Based on LDA, each observed word in the course description can be allocated to a certain topic and the course description is regarded as a mix of multiple topics. In this section, we first provide the related formulation, followed by an LDA-based model for course description clustering. Then, we propose the parameter inference process from the course description information.

3.1.1. Formulation. In our model, a collection of course description exists, $M = \{\mathbf{w}_m\}_{m=1}^{|M|}$ and $\mathbf{w}_m = \{w_{mi}\}_{i=1}^{N_m}$ is a vector of words in course description $\boldsymbol{m}$.

Definition 1. (Number definition). K, M, and V are the number of course topics, course descriptions, and unique words in all course descriptions, respectively. Words are indexed by $v \in \{1, 2, \cdots, V\}$, and N_m is the number of the word taken in course descriptions.

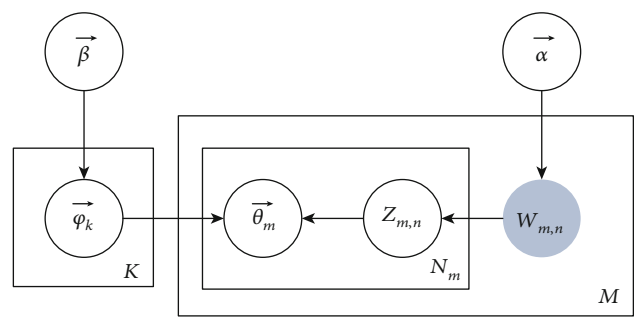


FIGURE 3: Graphical representation of LDA.

Definition 2. (Course topics and words). $Z_{m,n}$ is the topic associated with the n-th word in the course description $\boldsymbol{m}$, and $W_{m,n}$ is the n-th word in document $\boldsymbol{m}$.

Definition 3. (Variables for probability distribution). $\overrightarrow{\theta_m}$ is the multinomial distribution of topics specific to course description $\boldsymbol{m}$, which is a proportion for each course description, and each one is an $M * N$ matrix. $\overrightarrow{\varphi_k}$ is the multinomial distribution of words specific to the topics $\boldsymbol{k}$, which is a proportion for each topic and each one is a $K * V$ matrix.

Definition 4. (Variables for hyperparameter). $\vec{\alpha}$ is the hyperparameter to the multinomial distribution $\overrightarrow{\theta}$. $\vec{\beta}$ is the hyperparameter to the multinomial distribution $\vec{\varphi}$.

3.1.2. Model Description. This section presents the details of the LDA-based model for course description clustering. Figure 3 illustrates the relationships between the parameters used in the proposed model. The generative process is presented in Algorithm 1. For a better explanation, this model can be divided into two phases.

For each course topic $k \in [1, K]$:
(a) Draw a multinomial $\vec{\varphi}_k$ from a Dirichlet prior $\vec{\beta}$;
For each course description $m \in [1, M]$:
a. Draw a multinomial $\vec{\theta}_m$ from a Dirichlet prior $\vec{\alpha}$;
b. For each world $n \in [1, N_m]$ in course description m:
i. Draw a topic $Z_{m,n}$ from multinomial $\vec{\theta}_m$;
ii. Draw a word $W_{m,n}$ from multinomial $\vec{\varphi}_k(k = Z_{m,n})$;

Algorithm 1: Generative process of LDA.

(1) Phase 1: Modeling the Topic of the Course Description. In this model, we assume that each topic for the course description is represented by a word distribution. We model each topic $k \in \{1, 2, \cdots, K\}$ as a vector $\vec{\varphi}_k$ that follows a Dirichlet distribution over the V words.

$$\vec{\varphi}_k \sim \mathrm{Dir}(\beta), \tag{1}$$

where β is a symmetric Dirichlet prior.

(2) Phase 2: Modeling Words Distribution of Course Description. The key point of the LDA-based model for course description clustering is that each course description can be viewed as a mix of the latent topics, and each word in the course description has the corresponding topic. We model each course description $m \in \{1, 2, \cdots, M\}$ as a vector $\vec{\theta}_m$ that follows a Dirichlet distribution over the K topics.

$$\vec{\theta}_m \sim \mathrm{Dir}(\alpha), \tag{2}$$

where α is a symmetric Dirichlet prior.

We use the multinomial distribution $\overrightarrow{\theta_m}$ to sample a topic $Z_{m,n}$ for course contents. After determining the topic $Z_{m,n}$, we use the multinomial distribution $\overrightarrow{\varphi_k}$ to sample the word $W_{m,n}$.

3.1.3. Model Inference. The above process of the LDA-based model appears to be a relatively simple model but ensuring the accuracy of the derivation is difficult. We use Gibbs sampling to deal with this intractable question. Two steps (i.e., calculate the joint distribution and obtain the conditional distribution probability) are used to infer the parameters of the proposed model. The details of the reference process are as follows:

Calculate the Joint Distribution. The calculation of the joint distribution $P(\vec{W}, \vec{Z} \mid \vec{\alpha}, \vec{\beta})$ can be divided into two parts by

$$P\left(\vec{W}, \vec{Z} \mid \vec{\alpha}, \vec{\beta}\right) = P\left(\vec{W}, \vec{Z} \mid \vec{\beta}\right) P\left(\vec{Z} \mid \vec{\alpha}\right), \tag{3}$$

where $P(\vec{W}, \vec{Z} \mid \vec{\beta})$ is the probability of word generation in the entire course descriptions and $P(\vec{Z} \mid \vec{\alpha})$ is the probability of topic. Because the process of generating topics for the $\boldsymbol{M}$ courses in the course description sets is independent of each other, we can take the advantage of Dirichlet—the multinomial conjugated structure and conjugate priors to calculate the first probability in Equation (1) by

$$\begin{aligned} P\left(\vec{W}, \vec{Z} \mid \vec{\beta}\right) &= \int P\left(\vec{W} \mid \Phi, \vec{Z}\right) P\left(\Phi \mid \vec{\beta}\right) d\Phi \\ &= \int \prod_{\mathrm{k}=1}^{K} \prod_{\mathrm{v}=1}^{V} \varphi_{k,v}^{n_k^{(\mathrm{v})}} \prod_{k=1}^{K} \left(\frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \varphi_{k,v}^{\beta_v - 1} \right) d\Phi, \end{aligned} \tag{4}$$

where $n_k^{(v)}$ is the number of words $\boldsymbol{v}$ assigned to topic $\boldsymbol{k}$ and $\Gamma(x)$ in Equation (4) is the gamma function. In a similar way, $P(\vec{Z} \mid \vec{\alpha})$ can be calculated by

$$\begin{aligned} P\left(\vec{Z} \mid \vec{\alpha}\right) &= \int P\left(\vec{Z} \mid \theta\right) P\left(\theta \mid \vec{\alpha}\right) d\theta \\ &= \int \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{\mathrm{m,k}}^{n_{\mathrm{m}}^{(\mathrm{k})}} \prod_{\mathrm{m}=1}^{M} \left(\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{\mathrm{m,k}}^{\alpha_{\mathrm{k}} - 1} \right) d\theta, \end{aligned} \tag{5}$$

where $n_m^{(k)}$ represents the number of words in course description $\boldsymbol{m}$ assigned to topic $\boldsymbol{k}$.

Through Equations (4) and (5), we can obtain the joint contribution:

$$\begin{aligned} P\left(\vec{W}, \vec{Z} \mid \vec{\alpha}, \vec{\beta}\right) = &\left(\frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^{K} \left(\frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^{M} \\ &\times \prod_{\mathrm{k}=1}^{K} \frac{\prod_{v=1}^{V} \Gamma\left(n_k^{(v)} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} \left(n_k^{(v)} + \beta_v\right)\right)} \\ &\cdot \prod_{m=1}^{M} \frac{\prod_{k=1}^{K} \Gamma\left(n_m^{(k)} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} \left(n_m^{(k)} + \alpha_k\right)\right)}, \end{aligned} \tag{6}$$

Obtain the Conditional Distribution Probability. Using the chain rule, the conditional probability can be obtained as

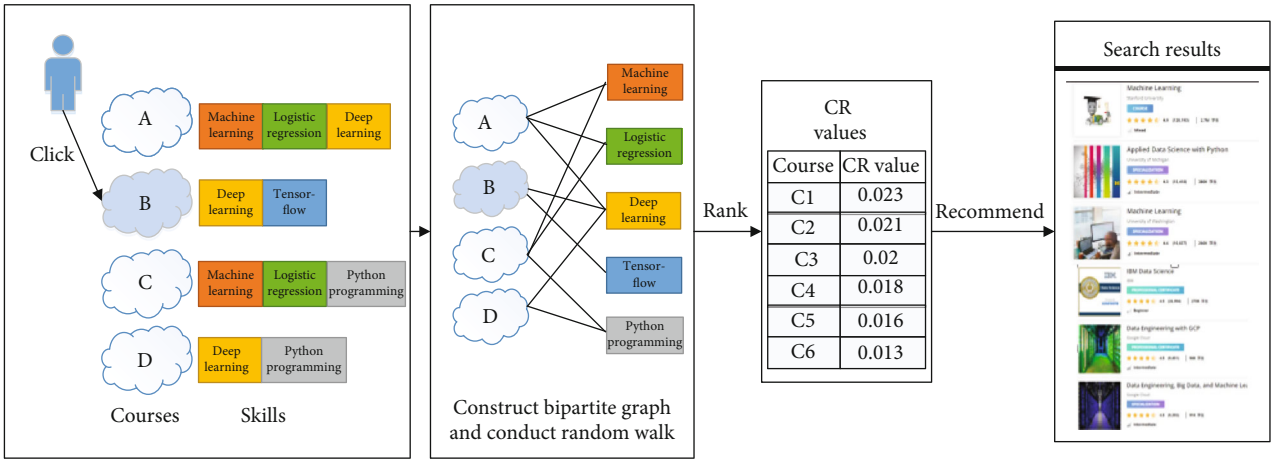


Figure 4: Framework of CourseRank algorithm.

$$P\left(Z_{m,n} \mid \vec{W}, \vec{Z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}\right) = \frac{P\left(Z_{m,n}, W_{m,n} \mid \vec{W}_{\neg(m,n)}, \vec{Z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}\right)}{P\left(W_{m,n} \mid \vec{W}_{\neg(m,n)}, \vec{Z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}\right)},$$

$$\propto \frac{P\left(\vec{W}, \vec{Z} \mid \vec{\alpha}, \vec{\beta}\right)}{P\left(\vec{W}_{\neg(m,n)}, \vec{Z}_{\neg(m,n)} \mid \vec{\alpha}, \vec{\beta}\right)} \propto \frac{n_{Z_{m,n}}^{(W_{m,n})} + \beta_{W_{m,n}} - 1}{\sum_{v=1}^{V}\left(n_{Z_{m,n}}^{(v)} + \beta_v\right) - 1} \times \left(n_m^{(W_{m,n})} + \alpha_{Z_{m,n}} - 1\right), \tag{7}$$

where $\neg(m, n)$ is a two-dimensional subscript, $\vec{W}_{\neg(m,n)}$ corresponds to all the words in the course descriptions except for the n-th word in the course description $\boldsymbol{m}$, $\vec{Z}_{\neg(m,n)}$ is the topic assignments for all words except for the n-th word in the course description $\boldsymbol{m}$.

Finally, based on the definition of Dirichlet–multinomial conjugated structure and Bayes rule, we can to obtain the multinomial parameter sets θ and $\Phi$by

$$P\left(\vec{\theta}_m \mid \vec{Z}_{m,n}, \vec{\alpha}\right) = \frac{P\left(\vec{\theta}_m \mid \vec{Z}_{m,n}, \vec{\alpha}\right)}{P\left(\vec{Z}_{m,n}, \vec{\alpha}\right)} = \frac{1}{Z_{\vec{\theta}_m}} \prod_{k=1}^{K} \theta_{m,k}^{n_m^{(k)} + \alpha_k - 1}$$
$$= \text{Dirichlet}\left(\vec{\theta}_m \mid \vec{n}_m + \alpha\right), \tag{8}$$

$$P\left(\vec{\varphi}_k \mid \vec{Z}, \vec{W}, \vec{\alpha}\right) = \frac{P\left(\vec{\varphi}_k, \vec{W} \mid \vec{Z}, \vec{\beta}\right)}{P\left(\vec{W} \mid \vec{Z}, \vec{\beta}\right)} = \frac{1}{Z_{\vec{\varphi}_k}} \prod_{v=1}^{V} \varphi_{k,v}^{n_k^{(v)} + \beta_v - 1}$$
$$= \text{Dirichlet}\left(\vec{\varphi}_k \mid \vec{n}_k + \vec{\beta}\right), \tag{9}$$

where $\vec{Z}_{m,n}$ is the topic assignments for all words in course description $\boldsymbol{m}$, that is, $\vec{Z}_{m,n} = \{Z_{m,n}\}_{n=1}^{N_m}$, $\vec{n}_m = \{n_m^{(k)}\}_{k=1}^{K}$ is the vector of topic observation counts for course description $\boldsymbol{m}$ and $\vec{n}_k = \{n_k^{(v)}\}_{v=1}^{V}$ that of word observation counts for topic $\boldsymbol{k}$. Using the expectation of the Dirichlet distribution on Equations (8) and (9), we can obtain the following result:

$$\varphi_{k,v} = \frac{n_k^{(v)} + \beta_v}{\sum_{v=1}^{V}\left(n_k^{(v)} + \beta_v\right)}, \tag{10}$$

$$\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K}\left(n_m^{(k)} + \alpha_k\right)}. \tag{11}$$

In Equations (10) and (11), $\varphi_{k,v}$and $\theta_{m,k}$ represent, respectively, the probability distribution of the words in content topic k and the probability of the content topics in course description m. From the perspective of MOOC recommendation, a topic may correspond to a knowledge point or a specific skill taught in various courses. We can consider each topic as a cluster and assign a course to the topic that corresponds to the largest course-topic probability in $\theta_{m,k}$. Based on $\theta_{m,k}$, we can also employ a classical algorithm to cluster the coursers.

3.2. Stage 2: Course Ranking Algorithm for MOOCs. With the clustering step in Stage 1, irrelevant courses can be filtered out for specific study purposes. However, many courses in each cluster remain, which would have a negative effect on the search ranking task for courses. Hence, to choose the right courses from a course cluster and present a precise ranking list for students, this paper designs an algorithm called CourseRank based on skills, which will gain through the courses to rank the courses in the same cluster.

The algorithm framework is illustrated in Figure 4. The figure shows that based on course skills, we construct a bipartite graph to rank the courses in the same clusters. The constructed bipartite graph consists of two kinds of disjoint and independent sets. The nodes on the left side represent courses and the nodes on the right side are skills. Based on the course-skill bipartite graph, we design the CourseRank algorithm to rank courses in each cluster when a student

Input: Bipartite graph G,ε,*root, maxstep*
Output: *CR* value
Initiative the root node $CR(root)$=1 and other nodes *CR* value is 0
while k < maxstep:
Set all nodes temp value are 0
From G, get node j and $j^{'}s$ out-edges set out_j
From out_j, get the nodes i connected to node j
compute relevance score: $temp[i] + = \varepsilon^{*} CR[j]/(len(\text{out}_j))$
temp[*roof*]+=(1-ε)
CR=*temp*
return *CR*

Algorithm 2: CourseRank algorithm.

searches for a keyword or clicks on a specific course. The proposed CourseRank algorithm, which is a variation of PageRank, is a strategy to rank nodes in a graph. In the PageRank algorithm, nodes are assumed to be connected with each other. However, this assumption cannot apply to the course-skill bipartite graph because we are required to estimate the relevance of all the courses to a specific course. Hence, we employ Equation (12) to compute the random access probability of a course node in CourseRank:

$$\text{PR}(i) = (1-\varepsilon)r_i + \varepsilon \sum_{j \in \text{in}(i)} \frac{\text{PR}(j)}{|\text{out}(j)|}. \tag{12}$$

In Equation (12), $\text{PR}(i)$ represents the probability that course i is accessed, $\text{in}(i)$ refers to all courses pointing to course i, and $\text{out}(j)$ represents other courses set up by course j. We replace $(1-\varepsilon)/N$ in classical PageRank algorithm with $(1-\varepsilon)r_i$ to compute the probability that course i will stay on the current course after being clicked on by the student as the starting point. Indicator r_i is 1 if the course is the target course and 0 otherwise. Equation (12) makes sure that, by walking randomly from the target course, the proposed CourseRank algorithm can compute the correlation from all other courses to the target course.

The algorithm details of CourseRank are presented in Algorithm 2.

The CourseRank algorithm will converge quickly to a stable state by calculating and updating the probabilities recursively. Based on CourseRank results, $\text{CR}(i)$ is used as the value to rank course i. We present the *top- k* courses in the same cluster of target courses or the *top- k* courses in the clusters associated with the search keywords.

4. Experiments

4.1. Dataset. The data used in our experiment were obtained from http://Coursera.com/, one of the most famous MOOC platforms in the world. Our data consisted of 2399 courses and 3981 course skills. The information related to each course included the course name, course description text in "About this Course," and the skill tags in "Skills you will gain." Because each course corresponds to several skills and each skill may be used to mark multiple MOOCs, the number of distinct skills in our data is 1590.

With the raw data obtained from the MOOC platform, we conduct the following preprocessing operations to obtain clean data:

(1) Convert all letters into lowercase and remove punctuation and meaningless words. After the preprocessing operation, the average length of the course descriptions is 90.79. The maximum length is 844 and the minimum length is 9. (2) Generate a word frequency matrix. In our experiment, we consider a course description as a document and the descriptions for all courses as the corpus. We construct a dictionary for the course corpus, assign a unique number to each word, and count the frequency of each word in the corpus. Because many course descriptions have words not related strongly to the course, we also conduct an operation to remove the noisy words from the corpus (e.g., a, able, about, and above). In our experiment, we have 19,746 distinct words in the course corpus

4.2. Course Clustering. We now evaluate the performance of the proposed LDA-based method to cluster courses.

4.2.1. Baseline Methods and Evaluation Metrics. In our paper, we designed an LDA-based method to cluster courses in MOOC platforms. In practice, many methods can group courses into clusters. For example, K-means [26] and DBSCAN [27] are the well-known clustering methods and are widely used for MOOC research. Chang et al. employed k-means to investigate the effects of learning style preferences on student intentions regarding MOOCs [28]. Chen et al. applied DBSCAN to cluster the learners into interested groups and analyzed their learning patterns of the groups [29]. This paper compares the proposed LDA-based method with k-means and DBSCAN. Before utilizing k-means and DBSCAN to cluster course descriptions, we use the TF-IDF method [30] to transform each course description as a numerical vector and conduct clustering with the TF-IDF matrices.

We use the coherence score to evaluate the performances of the proposed clustering model and k-means. Coherence

Table 1: Cluster results for courses.

Cluster	Typical courses
Cluster 1	(1) Advanced Instructional Strategies in the Virtual Classroom, (2) Blended Learning: Personalizing Education for Students, (3) Critical Issues in Urban Education, (4) Emerging Trends & Technologies in the Virtual K-12 Classroom, (5) Foundations of Teaching for Learning: Being a Teacher, (6) Foundations of Virtual Instruction, (7) Get Interactive: Practical Teaching with Technology, (8) Learning to Teach Online, (9) Powerful Tools for Teaching and Learning: Web 2.0 Tools, (10) University Teaching.
Cluster 5	(1) A Crash Course in Data Science, (2) Applied Plotting, Charting & Data Representation in Python, (3) Applying Machine Learning to your Data with GCP, (4) Basic Data Processing and Visualization, (5) Big Data Applications: Real-Time Streaming, (6) Building Data Visualization Tools, (7) Business Intelligence Concepts, Tools, and Applications, (8) Business intelligence and data warehousing, (9) Data Manipulation at Scale: Systems and Algorithms, (10) Foundations of marketing analytics.
Cluster 13	(1) Advanced Business Strategy, (2) Advanced Competitive Strategy, (3) Becoming a changemaker: Introduction to Social Innovation, (4) Business Growth Strategy, (5) Creating and Developing a Tech Startup, (6) Decision-Making and Scenarios, (7) Design Thinking for Innovation, (8) Design Thinking for the Greater Good: Innovation in the Social Sector, (9) Entrepreneurship 1: Developing the Opportunity, (10) FinTech Foundations and Overview.
Cluster 17	(1) Advanced Data Structures in Java, (2) Advanced R Programming, (3) Algorithmic Thinking, (4) An Introduction to Interactive Programming in Python, (5) Big Data Analysis with Scala and Spark, (6) Building Web Applications in PHP, (7) Cloud Computing Concepts, (8) Code Yourself! An Introduction to Programming, (9) Computational Thinking for Problem Solving, (10) Computer Science: Programming with a Purpose.
Cluster 26	(1) Adapt your leadership style, (2) Applications of Everyday Leadership, (3) Bridging the Gap between Strategy Design and Delivery, (4) Building High-Performing Teams, (5) Building Your Leadership Skills, (6) Designing and Implementing Your Coaching Strategy, (7) Giving Helpful Feedback, (8) Global sustainability and corporate social responsibility: Be sustainable, (9) Human Resources Management Capstone: HR for People Managers, (10) Influencing People.

score [31] is widely used to evaluate clustering quality. In our experiment, a course cluster is reasonable if the most probable words in the cluster cooccur more frequently in the course corpus. The coherence score is defined as follows:

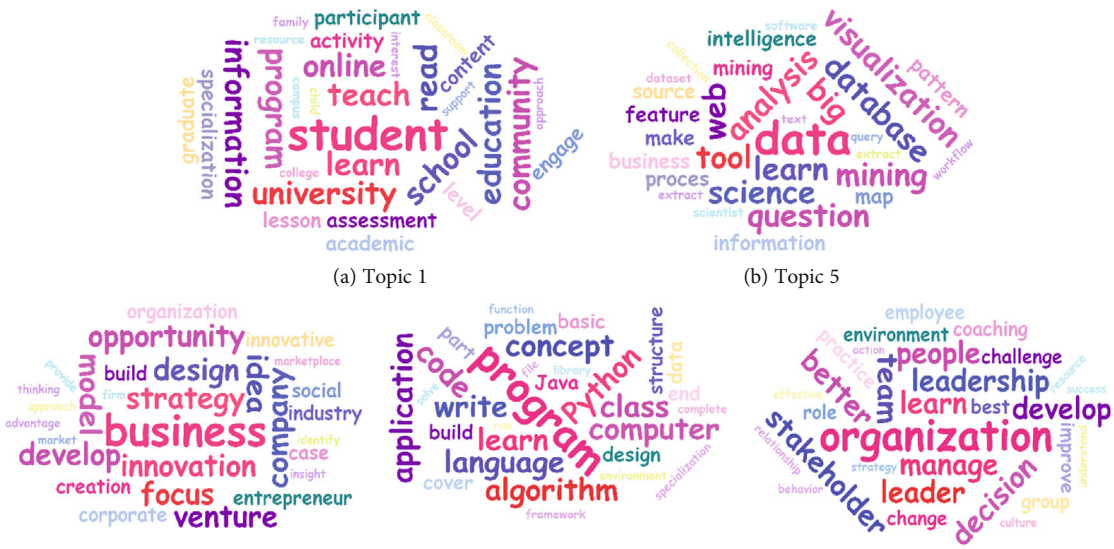
$$C\left(k\,;V^{(k)}\right)=\sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{D\left(v_m^{(k)},v_l^{(k)}\right)+1}{D\left(v_l^{(k)}\right)},\qquad(13)$$

where $V^{(k)}=(v_1^{(k)},\cdots,v_m^{(k)},\cdots,v_M^{(k)})$ is the list of the M most probable words in course cluster k, $D(v_l^{(k)})$ is the number of course descriptions containing word l, and $D(v_m^{(k)},v_l^{(k)})$ is the number of course descriptions containing word m and word l simultaneously.

4.2.2. Clustering Results. To obtain stable solutions, we run Gibbs samplers for 1000 iterations. In our experiment, $\alpha=50/K$ and $\beta=0.1$ where K is the number of clusters assumed by LDA. Based on the evaluation of optimal coherence value [32], both the number of clusters for the proposed method is set to be 36. To make a fair comparison, we predetermine the same cluster number for k-means.

We selected five clusters from the obtained clusters as examples and list them in Table 1. From Table 1, the proposed model can cluster courses with similar teaching objectives effectively. In Table 1, cluster 1 is a course group on teaching methodology. It gathers the courses for the new trend of teaching methods that can facilitate more effective learning environments. Students will gain skills, such as how to construct blended learning and how to organize interaction in the virtual classroom. Cluster 4 is a course group on data sciences, which includes content on data analysis, processing, visualization, and application in business intelligence and marketing. In the proposed model, course descriptions are analyzed by the LDA model. Therefore, we cluster courses according to their content topics rather than descriptive words. Many courses in the same clusters have distinct names but have similar teaching objectives for this reason. Cluster 12 is a course group on business strategy. In the cluster, we can see the courses that teach students how to formulate and innovate business strategies, especially in the new environment, such as the social and FinTech context. Cluster 16 contains courses about programming. Students can develop skills in data structure, programming language, and computational thinking ability. Based on the courses in cluster 25, students gain knowledge on how to build a team and form leadership in a team. From the courses, students can also learn how to communicate with others and optimize human resources management. In Figure 5, we illustrate the word clouds of the five clusters from which we can understand thoroughly the teaching objectives of the courses in each cluster.

Table 2 shows the comparison results on the coherence index between the proposed model and the baseline algorithms. We select the Top T words in each topic to evaluate the performance of these two methods. Table 2 shows that the proposed model always obtains the smaller coherence value regardless of the number of Top T. The proposed model performs better than k-means and DBSCAN. To test the robustness of the proposed method, we randomly split our data into two equal portions and cluster the courses in each portion by the three clustering methods. We illustrate the corresponding coherence scores on the top T representative words in Figure 6. From Figure 6, we can see that the



(a) Topic 1 (b) Topic 5

(c) Topic 13 (d) Topic 17 (e) Topic 29

Figure 5: Representative word clouds of clusters.

Table 2: Average coherence score on the top T representative words.

T	DBSCAN	k-means	LDA
5	-19.18	-17.14	-16.46
10	-115.46	-108.96	-92.18
15	-282.87	-262.51	-243.70
20	-541.50	-503.45	-472.54

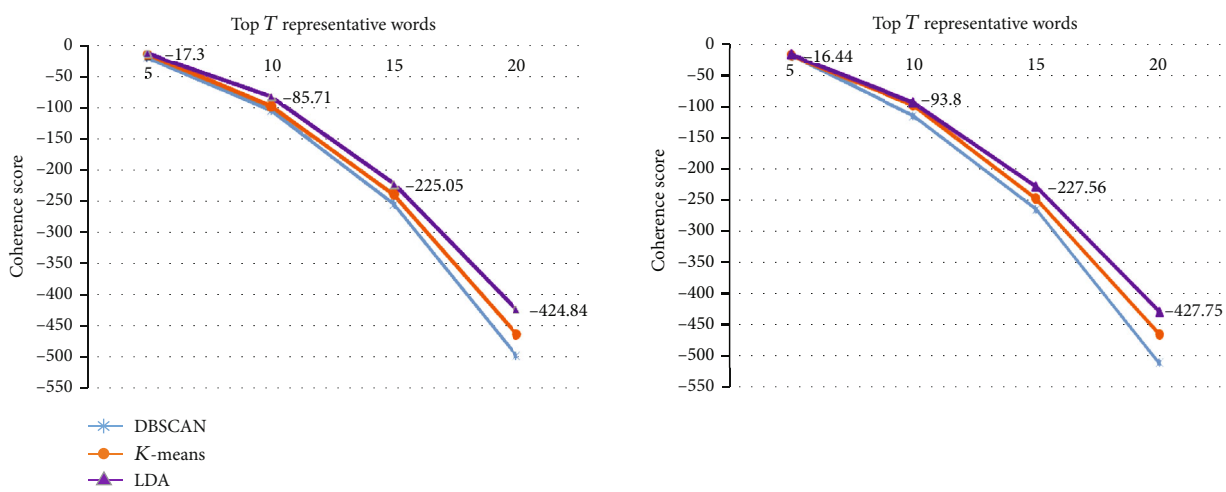
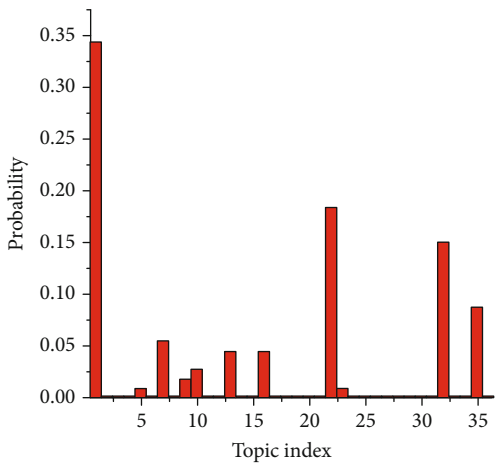


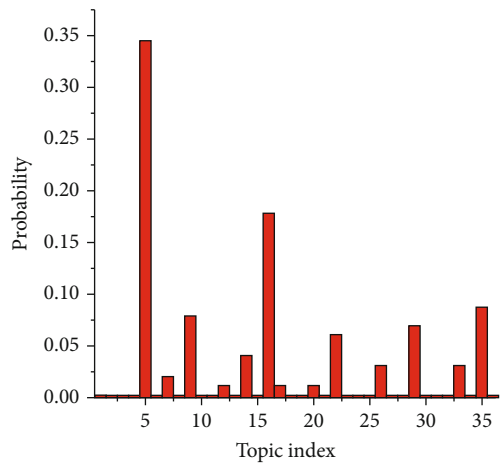
Figure 6: Robustness test experiment.

proposed LDA-based method is robust on the two datasets. And the LDA-based method always obtains smaller coherence values and performs better than k-means and DBSCAN.

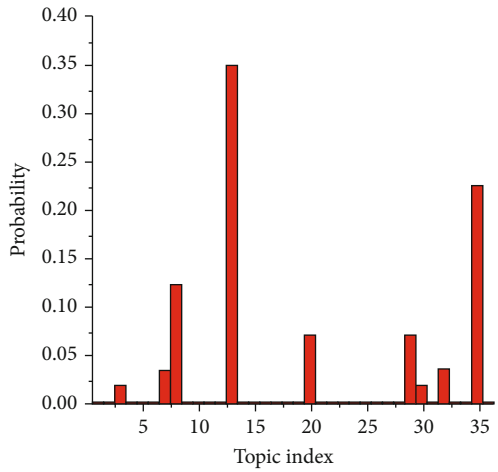
In the proposed model, we assign courses to clusters (topics) according to the course-topic distribution. A course is classified to the cluster corresponding to the maximum course-topic probability. We select one course from each cluster in Table 1 and illustrate their course-topic distribution in Figure 7. Figure 7 shows that the course–topic probabilities of the five courses are concentrated generally on one topic. For example, the probability of course "Building High-Performing Teams" belonging to topic (cluster) 29 is close to 50%, which is much bigger than the probabilities to other topics (clusters). Figure 7 indicates that the proposed LDA-based strategy can assign courses to the right clusters, which have a positive effect on the clustering results. In addition, clustering results in the k-means algorithm are affected significantly by the high-frequency words. The roles of the



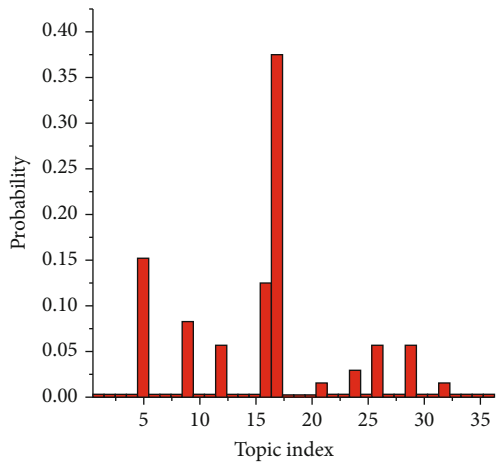
(a) Course Name: Blended Learning: Personalizing Education for Students



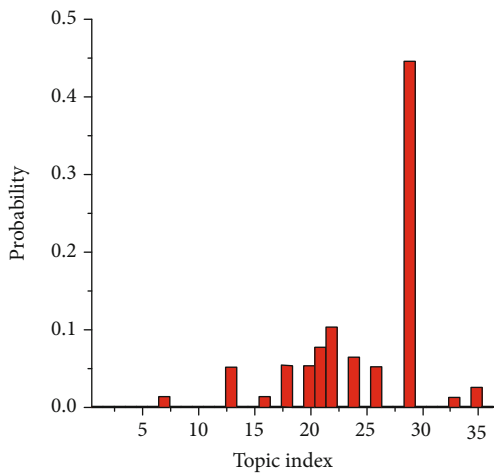
(b) Course Name: A Crash Course in Data Science



(c) Course Name: Advanced Business Strategy



(d) Course Name: Advanced R Programming



(e) Course Name: Building High-Performance Teams

Figure 7: Course-topic probability distribution.

nonhigh-frequency words, which indicate the teaching objectives, are likely to be weakened by the high-frequency words. In the proposed model, the courses are clustered according to topics rather than words. The latent topic strategy can smoothen the effects of the high-frequency words into multiple topics, thereby enabling us to obtain better clustering results than k-means.

4.3. Results on Course Ranking. Based on the course-topic (cluster) and the topic (cluster)-keyword distributions, we

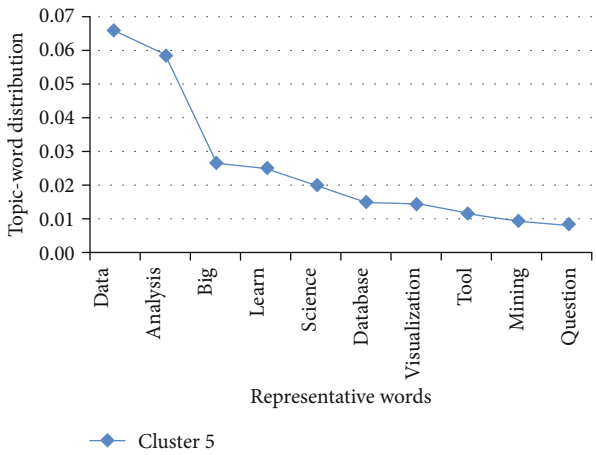


Figure 8: Representative words in Cluster 5 and their distribution probability.

Table 3: CourseRank values related to "Advanced Business Strategy.".

ID	Course name	CR value
1	Business growth strategy	0.029
2	Strategy formulation	0.027
3	Strategic management	0.026
4	Strategic organization design	0.021
5	Foundations of business strategy	0.019
6	Strategic planning and execution	0.017
7	Innovation and emerging technology: be disruptive	0.0166
8	(Re)-invent your business model with the odyssey 3.14 approach	0.015
9	Grow your business with Goldman Sachs 10,000 women	0.0145
10	Strategy implementation	0.0146

can optimize the course ranking task when students click on a specific course or search for a keyword through the search engine of a MOOC platform. If a student searches for a keyword through a search engine, we can filter out the topics unrelated to the keyword and list the courses in the related clusters according to the topic-keyword distribution. For example, if a student searches "Big data analysis," we can easily lock Cluster 5 as the target course cluster because its representative words are obviously related to "Big data analysis" (Figure 8). After locking the cluster, we can then show the representative courses in the cluster in the search list. In our experiment, courses such as "A Crash Course in Data Science" and "Applied Plotting, Charting & Data Representation in Python" belonging to Cluster 5 in Table 1 would be presented in the search list.

If a student clicks on a specific course in a MOOC platform, our experiment employs the CourseRank algorithm to rank the courses in the cluster where the clicked course belongs. For example, if a student clicks the course "Advanced Business Strategy," we employ the CourseRank algorithm to calculate the CourseRank value for each course in Cluster 13. The results are provided in Table 3. From Table 3, 10 courses are related to "Advanced Business Strategy," which is ranked in descending order by CourseRank values. These 10 courses together with the other courses would be shown in the recommendation lists of students who click on "Advanced Business Strategy." Similarly, if a student clicks on the course "Advanced Data Structures in Java" in Cluster 17, the proposed model would recommend the courses listed in Table 4 to the student.

Table 4: CourseRank values related to "Advanced Data Structures in Java.".

ID	Course name	CR value
1	Algorithms, part II	0.045
2	C++ for C programmers, part A	0.036
3	Algorithmic thinking (part 1)	0.031
4	C++ for C programmers, part B	0.015
5	Cloud computing concepts, part 1	0.0052
6	Algorithms, part I	0.0045
7	Data structures and performance	0.0042
8	Java programming: arrays, lists, and structured data	0.0035
9	Algorithmic thinking (part 2)	0.0034
10	Greedy algorithms, minimum spanning trees, and dynamic programming	0.0033

5. Conclusions

This paper proposed a novel search ranking method for MOOCs with the unstructured course descriptions and skills. The proposed model segments courses in the MOOC platforms into clusters based on course descriptions and ranks the courses in each cluster using course tags. This paper contributes theoretically to the educational research because we have introduced machine learning methods and employed new unstructured course information to deal with an important topic in the field.

Our experiments on the Coursera dataset showed that the proposed model can utilize the unstructured course description and skills efficiently to cluster courses and generate satisfactory search results. The experimental results indicated that the unstructured course descriptions and tags have rich information for MOOC services. Exploring the textual data using machine learning methods can help MOOC platforms improve recommendation accuracy. Figure 7 shows that a course usually provides knowledge across several education areas. Therefore, limiting a course to one education area would weaken service flexibility for MOOC platforms. The proposed models can help MOOC platforms position their courses accurately and improve their service qualities.

For future research, we will introduce more information to improve course ranking results. In this study, two kinds of unstructured data (i.e., course description and skills) were used to rank courses. In MOOC platforms, other kinds of data, such as word-of-mouth, can contain valuable information for the quality of courses. We will develop new search ranking models by considering these data. Another future

direction is to design methods to evaluate the search ranking results. Because we did not have the browsing logs of the search results, our experiment could not evaluate the accuracy of the obtained search ranking results. In the future, we will design subjective and objective strategies to test the effectiveness of the proposed method. The third direction is that many courses are missing learning skills in our study. Although we can infer the skills objectively from course contents and student reviews, new methods will be developed to infer course skills automatically.

Data Availability

Data are available for Requirement. Please send EMAIL to wqyao@ustc.edu.cn to obtain the data.

Ethical Approval

We have received approval from the ethics committee of Hefei University of Technology. We declare that no human participants were involved in this study.

Conflicts of Interest

We declare no conflict of interest concerning this study.

References

[1] R. F. Kizilcec, A. J. Saltarelli, J. Reich, and G. L. Cohen, "Closing global achievement gaps in moocs," *Science*, vol. 355, no. 6322, pp. 251-252, 2017.

[2] D. Shah, *By the Numbers: Moocs in 2018*, Class Central Moocreport, 2018, https://www.classcentral.com/report/mooc-stats-2018/.

[3] F. Bousbahi and H. Chorfi, "Mooc-rec: a case based recommender system for moocs," *Procedia-Social and Behavioral Sciences*, vol. 195, pp. 1813–1822, 2015.

[4] A. Elbadrawy and G. Karypis, "Domain-aware grade prediction and top-n course recommendation," in *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 183–190, Boston, MA, USA, 2016.

[5] M. Lin and D. W. Cheung, "An automatic approach for tagging web services using machine learning techniques1," in *Presented at Web Intelligence*, vol. 14, pp. 99–118, IOS Press, 2016.

[6] B. K. Daniel, "Big data and data science: a critical review of issues for educational research," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 101–113, 2019.

[7] J. A. Ruipérez-Valiente, S. Halawa, R. Slama, and J. Reich, "Using multi-platform learning analytics to compare regional and global mooc learning in the Arab world," *Computers & Education*, vol. 146, p. 103776, 2020.

[8] R. Bodily, H. Leary, and R. E. West, "Research trends in instructional design and technology journals," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 64–79, 2019.

[9] M. Zhu, A. Sari, and M. M. Lee, "A systematic review of research methods and topics of the empirical mooc literature (2014–2016)," *The Internet and Higher Education*, vol. 37, pp. 31–39, 2018.

[10] E. Costello, M. Brown, M. N. G. Mhichíl, and J. Zhang, "Big course small talk: twitter and moocs—a systematic review of research designs 2011–2017," *International Journal of Educational Technology in Higher Education*, vol. 15, no. 1, p. 44, 2018.

[11] M. Aparicio, T. Oliveira, F. Bacao, and M. Painho, "Gamification: a key determinant of massive open online course (mooc) success," *Information & Management*, vol. 56, no. 1, pp. 39–54, 2019.

[12] R. Deng, P. Benckendorff, and D. Gannaway, "Learner engagement in moocs: scale development and validation," *British Journal of Educational Technology*, vol. 51, no. 1, pp. 245–262, 2019.

[13] P. Luik, R. Suviste, M. Lepp et al., "What motivates enrolment in programming moocs?," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 153–165, 2019.

[14] Y. Sun, L. Ni, Y. Zhao, X. L. Shen, and N. Wang, "Understanding students' engagement in moocs: an integration of self-determination theory and theory of relationship quality," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 3156–3174, 2018.

[15] A. Cohen, U. Shimony, R. Nachmias, and T. Soffer, "Active learners' characterization in mooc forums and their generated knowledge," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 177–198, 2019.

[16] N. Hood, A. Littlejohn, and C. Milligan, "Context counts: how learners' contexts influence learning in a mooc," *Computers & Education*, vol. 91, pp. 83–91, 2015.

[17] P. J. Guo and K. Reinecke, "Demographic differences in how students navigate through moocs," in *Presented at Proceedings of the first ACM conference on Learning@ scale conference*, pp. 21–30, Atlanta, GA, USA, 2014.

[18] K. Li, "Mooc learners' demographics, self-regulated learning strategy, perceived learning and satisfaction: a structural equation modeling approach," *Computers & Education*, vol. 132, pp. 16–30, 2019.

[19] J. Reich and J. A. Ruipérez-Valiente, "The mooc pivot," *Science*, vol. 363, no. 6423, pp. 130-131, 2019.

[20] B. Hong, Z. Wei, and Y. Yang, "Discovering learning behavior patterns to predict dropout in mooc," in *Presented at 2017 12th International Conference on Computer Science and Education (ICCSE)*, pp. 700–704, Houston, TX, USA, 2017.

[21] D. Peng and G. Aggarwal, "Modeling mooc dropouts," *Entropy*, vol. 10, pp. 1–5, 2015.

[22] A. Ramesh, D. Goldwasser, B. Huang, H. Daume, and L. Getoor, "Understanding mooc discussion forums using seeded lda," in *Presented at Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pp. 28–33, Baltimore, Maryland USA, 2014.

[23] T. Atapattu and K. Falkner, "A framework for topic generation and labeling from mooc discussions," in *Presented at Proceedings of the Third (2016) ACM conference on learning@ scale*, pp. 201–204, Edinburgh, Scotland, UK, 2016.

[24] Y. Pang, Y. Jin, Y. Zhang, and T. Zhu, "Collaborative filtering recommendation for mooc application," *Computer Applications in Engineering Education*, vol. 25, no. 1, pp. 120–128, 2017.

[25] C. Lang, R. Levy-Cohen, C. Woo et al., "Automated extraction of learning goals and objectives from syllabi using lda and neural nets," *Presented at Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 2018.

[26] J. A. Hartigan and M. A. Wong, "Algorithm as 136: a k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, pp. 100–108, 1979.

[27] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Presented at Kdd*, vol. 96, pp. 226–231, 1996.

[28] R. I. Chang, Y. H. Hung, and C. F. Lin, "Survey of learning experiences and influence of learning style preferences on user intentions regarding MOOCs," *British Journal of Educational Technology*, vol. 46, no. 3, pp. 528–541, 2015.

[29] Y. Chen, Q. Chen, M. Zhao, S. Boyer, K. Veeramachaneni, and H. Qu, "Dropoutseer: visualizing learning patterns in massive open online courses for dropout reasoning and prediction," in *Presented at 2016 IEEE conference on visual analytics science and technology (VAST)*, pp. 111–120, Baltimore, MD, USA, 2016.

[30] R. R. Larson, "Introduction to information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 61, pp. 852-853, 2010.

[31] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Presented at Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, United Kingdom*, pp. 262–272, Edinburgh, Scotland, UK, 2011.

[32] S. Mankad, H. S. Han, J. Goh, and S. Gavirneni, "Understanding online hotel reviews through automated text analysis," *Service Science*, vol. 8, no. 2, pp. 124–138, 2016.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8843186, 8 pages
https://doi.org/10.1155/2020/8843186

Research Article

Multimodal Fusion Method Based on Self-Attention Mechanism

Hu Zhu,[1] Ze Wang,[2] Yu Shi,[3] Yingying Hua,[1] Guoxia Xu,[4] and Lizhen Deng[5]

[1]*Jiangsu Province Key Lab on Image Processing and Image Communication, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*
[2]*R&D Center, China Academy of Launch Vehicle Technology, Beijing 100176, China*
[3]*Bell Honors School, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*
[4]*Department of Computer Science, Norwegian University of Science and Technology, Gjovik 2815, Norway*
[5]*National Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*

Correspondence should be addressed to Lizhen Deng; alicedenglzh@gmail.com

Received 25 June 2020; Revised 10 August 2020; Accepted 2 September 2020; Published 23 September 2020

Academic Editor: Yin Zhang

Multimodal fusion is one of the popular research directions of multimodal research, and it is also an emerging research field of artificial intelligence. Multimodal fusion is aimed at taking advantage of the complementarity of heterogeneous data and providing reliable classification for the model. Multimodal data fusion is to transform data from multiple single-mode representations to a compact multimodal representation. In previous multimodal data fusion studies, most of the research in this field used multimodal representations of tensors. As the input is converted into a tensor, the dimensions and computational complexity increase exponentially. In this paper, we propose a low-rank tensor multimodal fusion method with an attention mechanism, which improves efficiency and reduces computational complexity. We evaluate our model through three multimodal fusion tasks, which are based on a public data set: CMU-MOSI, IEMOCAP, and POM. Our model achieves a good performance while flexibly capturing the global and local connections. Compared with other multimodal fusions represented by tensors, experiments show that our model can achieve better results steadily under a series of attention mechanisms.

1. Introduction

Multimodal integration has become a popular research direction in the field of artificial intelligence by virtue of its outstanding performance in various applications. Multimodal research has performed well in speech recognition [1], emotion recognition [2, 3], emotion analysis [4], speaker feature analysis [5], and media description [6].

Multimodal fusion is an extremely important research direction and core technology in multimodal field research. Multimodal fusion is aimed at utilizing the complementary information present in multimodal data by combining multiple modalities. It is one of the challenges of multimodal fusion to extend fusion to multimodal while keeping the model and calculation complexity reasonable.

Previous research methods used feature concatenation to fuse different data. These methods [7, 8] take the feature of the input concatenated as input, and some methods [9] even remove the temporal correlation in the modalities. Although these methods have been integrated at the beginning, it is precisely because of this that the interaction within the modal is suppressed at the beginning, causing the modalities to lose its overall correlation or even temporal dependencies.

Some fusion methods [10, 11] use methods such as weighted average or majority voting to fuse modalities together, and these modalities have their own models in later stages. Each of these methods has an inevitable shortcoming. Since each model is modeled separately, the interaction of the modes is lost.

At present, the latest methods [12, 13] try to use tensor representation to model the interactions between modes to solve those shortcomings. The extremely high-dimensional tensor representation caused by various forms of outer products puts a lot of pressure on the calculation

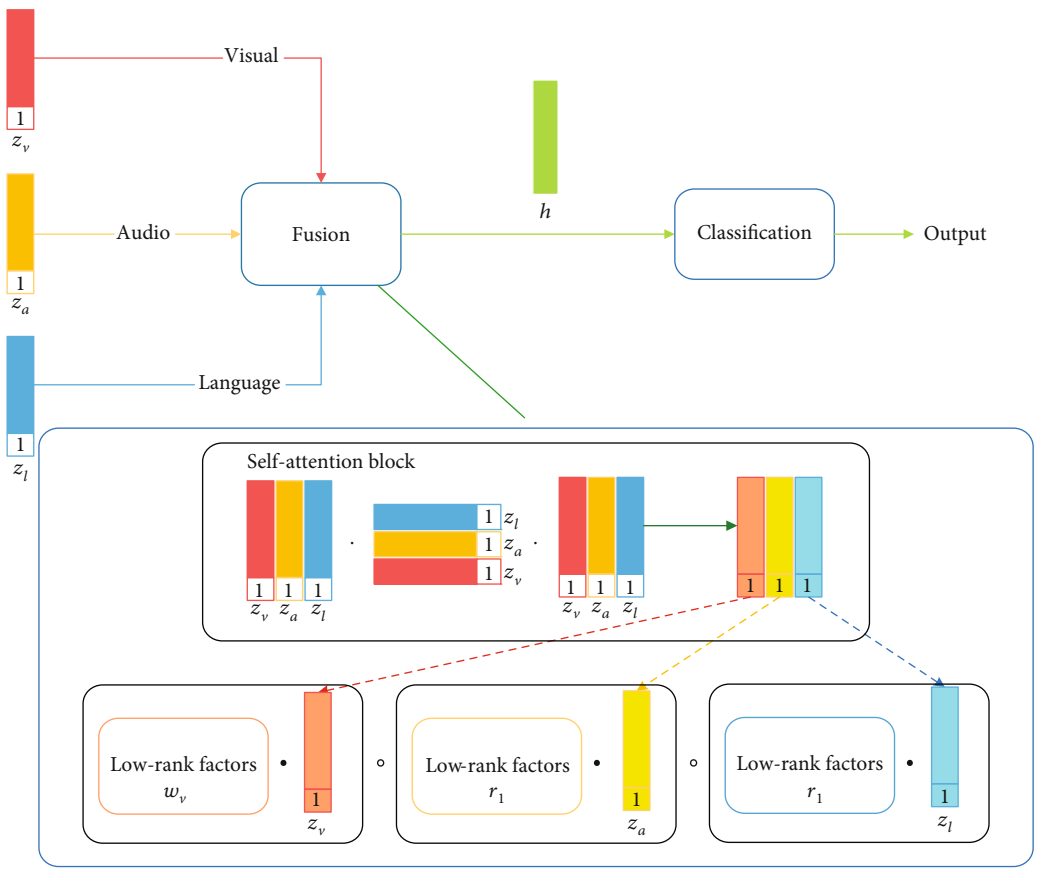


Figure 1: Overview of our multimodal fusion model based on self-attention mechanism: the unimodal representations z_v, z_a, and z_l as input to MF (multimodal fusion), which were obtained by passing the unimodal inputs x_v, x_a, and x_l into three subnetworks f_v, f_a, and f_l, respectively. In MF, z_v, z_a, and z_l generate new unimodal representations z'_v, z'_a, and z'_l through self-attention; then, z'_v, z'_a, and z'_l produce an output representation by performing low-rank multimodal fusion with modality-specific factors. The output will be multimodal representation, which can be used for applying classification task.

speed and memory occupation. In [14], Liu et al. proposed to use the low-rank multipeak fusion method, which partially solves the problem of large calculation and complicated parameters due to tensor representation but lacks the consideration of the correlation between multiple unimodal inputs.

An attention mechanism has been applied to various fields and has achieved satisfactory results. In [15], Wang et al. proposed "Residual Attention Network," a convolutional neural network using an attention mechanism which can incorporate with the state-of-art feed forward network architecture in an end-to-end training fashion. Lin et al. proposed a novel structure-attention-based LSTM as a hierarchical structure model, which has an advantage in capturing the potential semantic structure. As for applications, Choi et al. [16] proposed a fine-grained attention mechanism for neural machine translation while Ge et al. [17] proposed a leveraged attention mechanism in video action recognition. Hsiao and Chen [18] proposed to integrate the attention mechanism into deep recurrent neural network models for speech emotion recognition. However, none of these previous works aimed at applying an attention mechanism in multimodal fusion.

In this paper, we propose a novel low-rank multipeak fusion model based on a self-attention mechanism, which uses the low-rank weight tensor with an attention mechanism to make multipeak fusion more efficient and more globally relevant. The overall framework of our model is shown in Figure 1. We evaluate the performance of our method through experiments on three multimodal fusion tasks on public data sets and also compare our experiments with the latest models. While reducing the complexity and parameters of the model, we are studying how to improve the applicability and stability of our model. To our knowledge, this is the first time that the self-attention mechanism has been applied to the low-rank factor of multimodal fusion. Compared with other tensor-based models, our model performs very well both in terms of efficiency and performance.

The main contributions of our paper are as follows:

(i) We propose low-rank multimodal fusion based on a self-attention mechanism, which can effectively improve the global correlation

(ii) While maintaining low parameter complexity and high calculation speed, our model has high adaptability and can be applied to various tasks

(iii) We provide the performance of our model on three multimodal tasks evaluated on public data sets compared to other latest models

2. Related Work

2.1. Tensor Representation Method. The tensor representation method is one of the most successful methods for multimodal fusion. The core of tensor representation is to convert the input representation into a high-dimensional tensor, and then map it to a lower-dimensional output vector space. Tensors are usually formed by multiplying the outer product by the input modality. The input tensor Z is calculated from the unimodal representation:

$$Z = \otimes_{n=1}^{N} z_n, z_n \in \mathbb{R}^{d_n}, \tag{1}$$

where $\otimes_{n=1}^{N}$ denotes the tensor outer product over a set of vectors indexed by n, and z_n is the input representation.

The input tensor $Z \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ uses a linear layer $f(\bullet)$ to generate a vector representation:

$$h = f(Z\,; W\,; b) = W \cdot Z + b, h, b \in \mathbb{R}^{y}, \tag{2}$$

where W denotes the weight of this layer and b represents the bias. Because Z is an N-order tensor, where N is the number of input modes, the weight W should be an $N+1$-order tensor. The dimension of the weight W is $W \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N \times d_h}$, where the $N+1$th dimension is equal to the size of the output representation d_h. Since $W \cdot Z$ is a dot product, the weight W can be regarded as $d_h N$th order tensor.

Due to the high dimension of tensor Z, the computational difficulty and model complexity of tensor fusion method are greatly improved. The dimension of tensor Z increases exponentially with the number of modes. This makes the tensor fusion method fail to perform more tasks at the same time, which reduces the adaptability of the model.

2.2. Low-Rank Tensor Representation Method. The low-rank multimodal fusion method is aimed at solving the shortcomings of the multimodal fusion model represented by a tensor with the method of decomposing the weight W into a set of low-rank factors.

The method of degrading the weights in the multimodal fusion method represented by the low-rank tensor is to decompose the weight W into N fixed modalities. Because $W \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N \times d_h}$ can be regarded as $d_h N$th order tensor, so our weight can be expressed as follows:

$$\widetilde{W_m} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}, m = 1, \cdots, d_h. \tag{3}$$

A deep understanding of formula (3), so $\widetilde{W_m}$ has the following exact decomposition of the vector in Equation (5). Each $\widetilde{W_m}$ contributes to one dimension in the vector h, so we can simplify Equation (2):

$$h_m = \widetilde{W_m} \cdot Z, \tag{4}$$

$$\widetilde{W_m} = \sum_{i=1}^{R} \otimes_{n=1}^{N} w_{n,m}^{(i)}, w_{n,m}^{(i)} \in \mathbb{R}_n^d, \tag{5}$$

where R is the rank of the tensor, which makes the decomposition most efficient. $\{\{w_{n,m}^{(i)}\}_{n=1}^{N}\}_{i=1}^{R}$ is the decomposition factor of the original weight tensor based on rank R.

In the above formula $\{\{w_{n,m}^{(i)}\}_{n=1}^{N}\}_{i=1}^{R}$, we give the rank R a fixed value r, and the formula $\{\{w_{n,m}^{(i)}\}_{n=1}^{N}\}_{i=1}^{r}$ can be decomposed by the fixed rank, and the model is parameterized at the same time. We expand the vector $w_{n,m}^{(i)}$ by m (where $= 1, \cdots, d_h$) into a set of low-rank factors $w_n^{(i)} = [w_{n,1}^{(i)}, w_{n,2}^{(i)}, \cdots w_{n,dh}^{(i)}]$, so the $\{w_n^{(i)}\}_{i=1}^{r}$ is its corresponding low-rank factors. Therefore, the weights of the multimodal fusion method represented by tensor can be transformed into low-rank weight tensor:

$$W = \sum_{i=1}^{r} \otimes_{n=1}^{N} w_n^{(i)}. \tag{6}$$

Bring Equation (6) into Equation (2) to get the following a simplified low-rank tensor representation:

$$\begin{aligned} h = W \cdot Z &= \left(\sum_{i=1}^{r} \otimes_{n=1}^{N} w_n^{(i)} \right) \cdot \left(\otimes_{n=1}^{N} z_n \right) \\ &= \sum_{i=1}^{r} \left(\otimes_{n=1}^{N} w_n^{(i)} \cdot \otimes_{n=1}^{N} z_n \right) = \bigwedge_{n=1}^{N} \left(\sum_{i=1}^{r} w_n^{(i)} \cdot z_n \right). \end{aligned} \tag{7}$$

We made a series of derivation changes in the above formula and finally turned the model calculated from an exponentially complex model into a linear model, where $\bigwedge_{n=1}^{N} x_n$ denotes the product of elements in the order of tensors: $\bigwedge_{n=1}^{N} x_n = x_1 \circ x_2 \circ \cdots \circ x_N$.

Compared with the original tensor representation method, the low-rank multimodal fusion method improves the calculation speed and reduces the complexity of the model. However, only a simple outer product operation is performed for each single mode, which largely ignores the correlation between each single mode and loses the global uniformity.

2.3. Attention Mechanism. Neural networks equipped with attention have parallelizable computation, lightweight structure, and the ability to capture both long-range and local dependencies. The core of the attention mechanism method is to measure the correlation between z_n and q. A compatibility function $g(z_n, q)$ generates score k, which can reflect the dependency between z_n and q. The score is converted into a probability by function softmax, and finally, the probability is used as a weight.

$$k = [g(z_n, q)]_{n=1}^{N}, \tag{8}$$

$$p(y|z, q) = \text{softmax}(k), \tag{9}$$

$$s = \sum_{n=1}^{N} p(y = n|z, q) \cdot z_n, \tag{10}$$

where k is represented as a vector of n correlation scores. By applying k to the function softmax, we get a probability distribution about attention $p(y|z, q)$. And s is the output vector for query q.

In the attention mechanism, choosing different compatibility functions $g(z_n, q)$ will have different experimental results. The different compatibility functions also directly lead to various categories of attention mechanisms. In this paper, the attention mechanism of our method uses the dot product attention compatibility function as follows:

$$g(z_n, q) = \left\langle w^{d_1} z_n, w^{d_2} q \right\rangle, \tag{11}$$

where w^{d_1}, w^{d_2} are learnable parameters, $\langle \bullet, \bullet \rangle$ denotes the inner product.

3. Our Methods

3.1. Overview. The method proposed in this paper is an improvement to the low-rank multimodal fusion method and an effective improvement to the input modal based on the low-rank multimodal fusion method. We propose a novel self-attention mechanism and apply it between input modalities to improve the correlation and local dependence among various modalities. We pay more attention to the improvement of the self-mode, so we choose to use the self-attention mechanism instead of the traditional attention mechanism model. Since our model does not introduce redundant parameters, our model maintains a low complexity while improving accuracy. In addition, our self-attention module uses parallel computing, which makes the calculation speed greatly improved compared with the traditional attention mechanism model. Compared with the model using traditional attention mechanism, our model has lower complexity and faster running speed.

3.2. Network Architecture. The overall framework of our network model is shown in Figure 1. Our model network is composed of three parts, namely, the extraction module, fusion module, and classification module. The fusion module is the core part of our model, which is what we will focus on next. The task of the extraction module is to transform the unimodal inputs x_v, x_a, and x_l into unimodal representation s z_v, z_a, and z_l through the subnetworks f_v, f_a, and f_l. The unimodal representation obtained by the extraction module is expressed in the form of tensor, which is more convenient for the following calculation. And the fusion module contains a self-attention module for each unimodal representation. The unimodal representation enters the fusion module and generates a unimodal representation with new weights through a self-attention mechanism. Observing our network model, we do not need to directly calculate the input tensor Z, we first decompose z_v, z_a, and z_1 in low rank to get z_v, z_a, and z_1,then assign the corresponding weights W_v, W_a, W_1 to each factor, and finally sum them with the weights, which greatly reduces the complexity of our model and reduces the calculation pressure. Finally, the input tensor passing through the self-attention module generates the output tensor in the fusion module, which is the final output result that can be used for classification.

3.3. Self-Attention Module. Since each unimodal has different information, the purpose of multimodal fusion is to make full use of the complementary information of multimodal data. We note that the self-attention module also has the ability to capture the global and local connections, so the most prominent part of our contribution in this article is to propose the introduction of the self-attention module into multimodal fusion. In the self-attention module, we use a different output vector calculation method than the traditional attention mechanism. This new method can perfectly meet the requirements of our simultaneous input of multiple tasks and realize parallel computing. The self-attention model we proposed is a weighted self-attention in proportion, which includes multitask self-attention. Our self-attention model formula is as follows:

$$s = v\text{softmax}\left(\frac{q^T k}{\sqrt{d_q}}\right)^T, s = [s_1, s_2, \cdots, s_n] \in \mathbb{R}^{d_{i\times n}}. \tag{12}$$

The three parameters q, k, and v of Equation (12) all conform to this equation $q, k, v = \varphi^{q,k,v}(z_n)$, which means that all three input parameters come from the same source, where $v \in \mathbb{R}^{d_{i\times n}}, k \in \mathbb{R}^{d_{n\times i}}, q \in \mathbb{R}^{d_{i\times n}}$. For the multitask attention mechanism, the input will be projected into multiple subspaces. This parameter uniformly scales the dot product attention to be embedded in each subspace.

Since $s = [s_1, s_2, \cdots, s_n] \in \mathbb{R}^{d_{i\times n}}$ is a series of output vectors of q, k, and v, therefore, s is a series of output vectors of z_n, we derive the following equation:

$$z_n' = s_i, z_n' \in \mathbb{R}^{d_n}, \tag{13}$$

where z_n' is the new unimodal representation generated by the self-attention module. In this way, the self-attention between our single modes is completed. Bring Equation (13) into Equation (7) and simplify it as follows:

$$h = W \cdot Z = \left(\sum_{i=1}^{r} \otimes_{n=1}^{N} w_n^{(i)}\right) \cdot \left(\otimes_{n=1}^{N} z_n'\right) = \bigwedge_{n=1}^{N} \left(\sum_{i=1}^{r} w_n^{(i)} \cdot z_n'\right). \tag{14}$$

It can be seen from formula (14) that the formula is consistent with the model we have shown. First, superimpose each weighting factor, then do element product between each single module.

Since we are merging multiple tasks at the same time, we will show below that when $n = 2$, our formula will expand to formula (15):

$$\begin{aligned} h &= \bigwedge_{n=1}^{2}\left(\left(\sum_{i=1}^{r} w_1^{(i)} \otimes w_2^{(i)}\right) \cdot z_n'\right) \\ &= \left(\sum_{i=1}^{r} w_1^{(i)} \cdot z_1'\right) \circ \left(\sum_{i=1}^{r} w_2^{(i)} \cdot z_2'\right). \end{aligned} \tag{15}$$

In this way, we can appropriately expand the formula according to the actual situation. It can be seen that the proposed method has high adaptability and can be flexibly applied in various tasks. In our self-attention module, we can see that our new input representation represents multiple tasks applied to multimodal fusion. And our self-attention module uses parallel computing to improve the accuracy of the model while maintaining a high speed of model calculation.

3.4. Training Loss. Our model adopts the mean absolute error (MAE) as our loss function. MAE is the average value of absolute error, which can better reflect the actual situation of classification error and can also reflect the classification performance of our model.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| h_i - h_i^P \right|}{n}, \tag{16}$$

where h_i is the output tensor we got through our model and h_i^P is the classified value. n represents the total number of our training samples.

4. Experiment

4.1. Experimental Environment. Our tensor representation method is generally based on a tensor fusion network, but the biggest difference from this network is that our method uses a self-attention mechanism in the MF module. In the experiment, we compared our method with some of the latest multimodal fusion methods. Our experiment environment is 2080Ti*2 Graphic Processing Unit (GPU), 32 G memory, 12 Intel(R) Xeon(R) W-2133 CPU @ 3.60 GHz. Our model training and testing are completed on CONDA 4.8.3, python3.7.7, and pytorch 1.5.0.

4.2. Data sets. We conduct our experiments on multimodal data sets; they are CMU-MOSI [19], IEMOCAP [20], and POM [6]. These data sets provide data for sentiment analysis, speaker feature recognition, and emotion recognition. The goal of our experiment is to identify the speaker's emotions through these verbal or nonverbal behaviors.

4.2.1. IEMOCAP. This IEMOCAP data set is a collection of 151 recorded dialogue videos; each dialogue video has two speakers, so the entire data set has a total of 302 videos. Each video is marked with 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral).

4.2.2. POM. The POM data set consists of 903 movie review videos, and the speakers of each video are marked with confidence, enthusiasm, and other characteristics.

4.2.3. CMU-MOSI. 93 movie review videos on YouTube make up the data set CMU-MOSI. Each video contains multiple opinion segments, and each segment has tags about the opinion on sentiment.

4.3. Comparisons. We compare the results of our method with the following baselines and state-of-the-art models: support vector machines (SVM) [21], deep fusion model (DF) [11], bidirectional contextual LSTM (BC-LSTM) [13], multi-view LSTM (MVLSTM) [22], low-rank multimodal fusion network (LMF) [23], and hierarchical polynomial fusion network (HPFN) [24]. We report mean absolute error (MAE) and accuracy of classification, F1 score, and Pearson's correlation (Corr).

4.4. Evaluation Metrics. We report four evaluation metrics as used by our multiple task: F1-emotion, accuracy Acc-k where k is the number of classes, mean absolute error (MAE), and Pearson's correlation (Corr). Among those metrics, F1-emotion is the score of the model under different emotions; as a statistical measure of the accuracy of a binary classification model, it can be viewed as a weighted average of the model accuracy and recall with a maximum of 1and a minimum of 0. Accuracy (Acc) is defined as the percentage of the total sample that classifies the correct result. Mean absolute error (MAE) reflects the classification performance as we reported before. Pearson's correlation (Corr) considers the degree of correlation among variables.

4.5. Implementation Details. We use the CANDECOMP/-PARAFAC(CP) decomposition format as the tensor networks in our experiments as we mentioned in Equation (6). Following LMF, we choose the candidate CP ranks {1, 4, 8, 16}.The result in different ranks are reported in Figure 2.

4.6. Experimental Data Analysis. We compare our method's performance on the three tasks of sentiment analysis, speaker feature recognition, and sentiment recognition with the previous models with excellent performance. The results are shown in Table 1. In all data sets, our approach can produce competitive and consistent results across metrics such as F1, Corr, Acc, and MAE.

On the emotion recognition task, our model got the highest score on three emotions scored by F1. The results verify that our method outperforms other traditional method and is close to the state-of-the-art approaches.

On the multimodal personality trait recognition task, our model also achieved competitive results. Although LMF achieved a high score on the ACC indicator, our score is only 0.1 less than the LMF score.

On the multimodal sentiment analysis task, our model performs very well on performance indicators Corr and Acc-2. Nonetheless, our method scored only 0.057 less than the highest on MAE. All in all, our method perfectly completes the multimodal sentiment regression task.

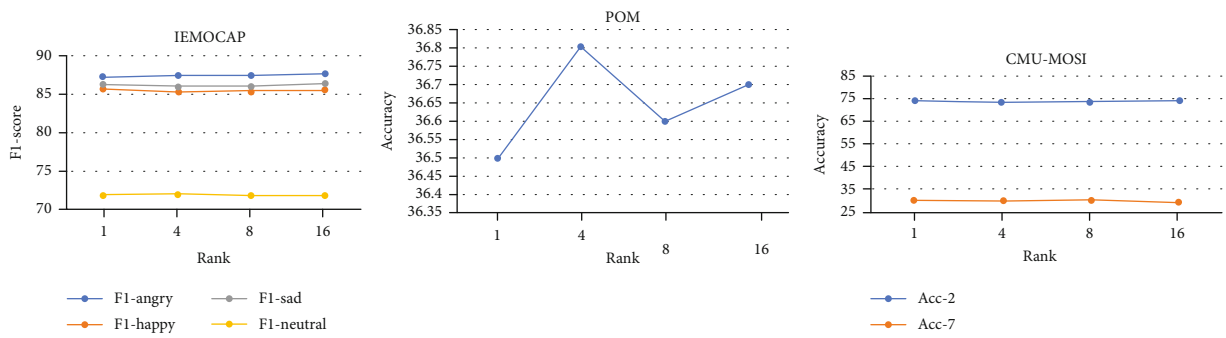


Figure 2: Results for recognition in different rank on IEMPCAP, POM, and CMU-MOSI.

Table 1: Results for emotion recognition on IEMPCAP, personality trait recognition on POM, and sentiment analysis on CMU-MOSI.

Data set	IEMOCAP				POM			CMU-MOSI				
Metric	F1-happy	F1-sad	F1-angry	F1-neutral	MAE	Corr	Acc	MAE	Corr	Acc-2	F1	Acc-7
SVM	81.5	78.8	82.4	64.9	0.887	10.4	33.9	1.864	0.057	50.2	50.1	17.5
DF	81.0	81.2	65.4	44.0	0.869	14.4	34.1	1.143	0.518	72.3	72.1	26.8
BC-LSTM	81.7	81.7	84.2	64.1	0.840	27.8	34.8	1.079	0.581	73.9	73.9	28.7
MV-LSTM	81.3	74.0	84.3	66.7	0.891	27.0	34.6	1.019	0.601	73.9	*74.0*	33.2
LMF	85.2	85.8	87.4	71.7	0.837	32.3	*36.8*	1.071	0.571	73.3	73.3	30.0
HPFN	*85.7*	86.2	*87.8*	*71.9*	0.840	35.6	36.7	*0.975*	0.601	73.0	73.1	*35.1*
Our methods	*85.7*	*86.3*	87.6	*71.9*	*0.834*	*52.3*	36.7	1.032	*0.610*	*74.1*	73.0	30.5

Best results are italicized.

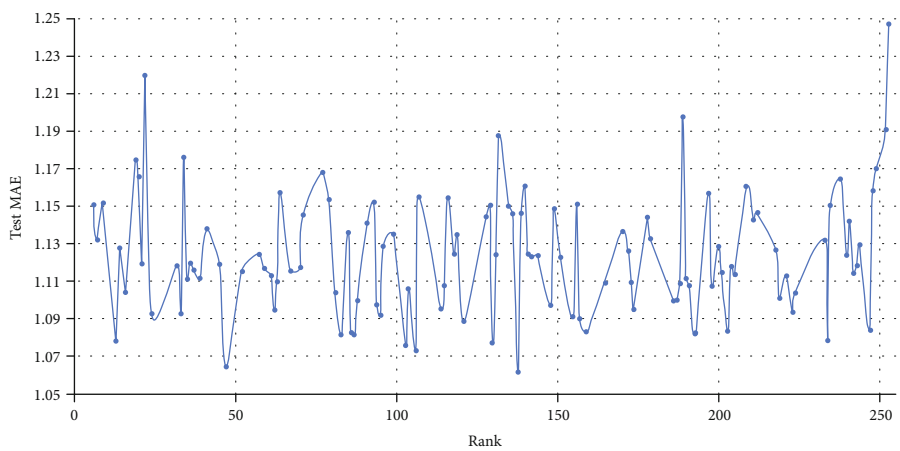


Figure 3: The effect of different level settings on the model performance: as the level increases, the results do not change significantly.

4.7. Influence of Rank Setting. In the experiment, the parameters changed by the actual situation often have a great influence on the experimental results. The different rank settings in our model will indeed affect the experimental results. In order to prove that our model can stand out in various tasks and has a high adaptability, we propose a new experiment, in which, we constantly set different values for rank to observe the effect of changes in rank on the experimental results.

In the experiment of the influence of rank on the experimental results, our other parameters are set as follows: the dropout of audio and video are both set to 0.2, and the text dropout is set to 0.5. For some other parameters, the learning rate is set to 0.001, batch size is set to 32, and the weight decay is set to 0.01.

To evaluate the impact of different level settings on our model, we measured the performance change of MAE in the CMU-MOSI data set while changing the number of levels. The results are shown in Figure 3. We have observed that although the rank value is constantly increasing, our training results have remained stable. Therefore, it can be seen that our model is not sensitive to rank, no matter what the rank is, the performance of our model can always remain stable. In some cases where the rank value is high, our model can still be adapted and used.

5. Conclusion

In this paper, we propose a multipeak fusion method based on a self-attention mechanism. This method uses a low-rank tensor representation, and the attention mechanism is used in tensor representation to improve the correlation between multiple representations. Our method achieves competitive results in different multimodal fusion tasks in different data sets. Our method reduces the complexity of the parameters while also reducing the measurement complexity. It is a novel attempt to apply the attention mechanism to multimodal fusion, and it shows higher efficiency and better performance on different downstream tasks. The application of the attention mechanism makes our model have higher classified ability under the premise of few parameters and high efficiency. In the experiment, our method performs better than the multimodal fusion method represented only by low-rank tensor.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61701259.

References

[1] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 1989.

[2] S. L. Chen, S. T. Huang, M. Tsutomu, and N. Ryohei, "Multimodal human emotion/expression recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, April 1998.

[3] W. Weninger and K. Schuller, "Youtube movie reviews: in, cross, and open-domain sentiment analysis in an audiovisual context," *Asbury Theological Seminary*, vol. 28, pp. 46–53, 2013.

[4] L. C. D. Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. No.97TH8237)*, Singapore, Singapore, September 1997.

[5] Y. Attabi and P. Dumouchel, "Anchor models and WCCN normalization for speaker trait classification," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, pp. 522–525, Portland, OR, USA, September 2012.

[6] S. Park, H. Shim, M. Chatterjee, K. Sagae, and L. Morency, "Computational analysis of persuasiveness in social multimedia," in *ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 50–57, New York, NY, USA, November 2014.

[7] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain, December 2016.

[8] H. Wang, A. Meghawat, L. Morency, and E. Xing, "Select-additive learning: improving cross-individual generalization in multimodal sentiment analysis," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, China, 2016.

[9] L. P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: harvesting opinions from the web," in *ICMI '11: Proceedings of the 13th international conference on multimodal interfaces*, New York, NY, USA, November 2011.

[10] T. Wortwein and S. Scherer, "What really matters — an information gain analysis of questions and reactions in automated PTSD screenings," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 15–20, San Antonio, TX, USA, October 2017.

[11] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrusaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *ICMI '16: Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, October 2016.

[12] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, http://arxiv.org/1606.01847.

[13] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, http://arxiv.org/1707.07250.

[14] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, and L. P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018.

[15] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6450–6458, Honolulu, HI, USA, 2017.

[16] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, 2018.

[17] H. Ge, Z. Yan, W. Yu, and L. Sun, "An attention mechanism based convolutional lstm network for video action recognition," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 20533–20556, 2019.

[18] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2526–2530, Calgary, AB, Canada, April 2018.

[19] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[20] C. Busso, M. Bulut, C.-C. Lee et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] S. S. Rajagopalan, L. P. Morency, T. Baltrušaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *European Conference on Computer Vision*, Amsterdam, Netherlands, 2016.

[23] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, http://arxiv.org/1806.00064.

[24] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8891393, 8 pages
https://doi.org/10.1155/2020/8891393

Research Article

A Visual Tracking Method Based on an Adaptive Overlapping Correlation Filter for Robotic Real-Time Cognitive Imaging

Yihua Lan, Pianpian Ma, Anfeng Xu, and Jinjiang Liu

School of Computer and Information Technology, Nanyang Normal University, Nanyang 473061, China

Correspondence should be addressed to Jinjiang Liu; jjliunynu@sina.com

Received 6 June 2020; Revised 13 July 2020; Accepted 31 July 2020; Published 24 August 2020

Academic Editor: Yin Zhang

Computer vision is a very important research direction in the cognitive computing field. Robots encounter various target-tracking problems with computer vision systems. Robust scale estimation is an important issue in tracking algorithms. Most of the available methods have difficulty addressing even reasonable changes of scale in complex videos. In this paper, we propose a visual tracking method based on robust scale estimation, which uses a discriminant correlation filter based on a time-dependent scale-space filter and an adaptive cross-correlation filter. The tracker uses separate essential filters for sample migration and scale estimation. Furthermore, the built-in scale estimation method can be introduced into other tracking algorithms. We validate the proposed method on the UAV123 dataset. The results of comparison experiments with the traditional correlation filter tracking method demonstrate that the proposed method improves the success rate and tracking accuracy while controlling the computational complexity; its success rate measured by the area under the curve is 0.638, while at a location error precision of 20%, it is 0.649.

1. Introduction

Computer vision is the basis of cognitive imaging and processing for robots. Robots interact with the external environment in real time through computer vision systems. After acquiring accurate control information, they can control themselves through a closed-loop process to achieve various tasks requiring artificial intelligence, including positioning, navigation, formation, cooperation, group intelligence, and even tasks requiring intelligent decisions and collaborative work. Therefore, the study of computer vision—especially real-time moving object tracking—is an important research field in robot cognitive computing. Through many unremitting scholastic efforts, real-time target-tracking technology has made considerable progress. However, due to the complexity of visual tracking systems and the variability of the targets themselves, robust target tracking is difficult.

Despite great progress in recent years, the target-tracking problem remains intractable, mainly due to partial occlusion interference, deformations, motion blur, illumination changes, and cluttered backgrounds in the video image sequences. In particular, scale changes are difficult to estimate during the tracking process. Similar problems also arise in environment identification [1], image reconstruction [2], and intelligent systems [3].

Research on visual tracking algorithms is ongoing and has made great progress in terms of stability and accuracy, but many problems remain to be solved. Because image processing involves the use of large quantities of data, to achieve the goals of better image recognition and tracking, it is necessary to implement more complex algorithms with more computations, which leads to the development of tracking algorithms with higher complexities. This higher algorithm complexity generally results in greater stability and accuracy but degrades real-time performance. Practical applications involving visual tracking usually have stringent real-time tracking and accuracy requirements, but the lag in hardware development makes it difficult for complex tracking algorithms to run in real time on embedded hardware. Therefore, image tracking must balance algorithm complexity with the available hardware computing power to achieve a satisfactory tracking effect.

In 2010, Bolme et al. [4] first introduced the correlation filter in signal processing into target tracking and proposed

the minimum output sum of squared error (MOSSE) filter method, whose discrimination model is based on the least square error, thereby creating a new tracking approach. The principle of this model is simple, it operates rapidly, and it is better at distinguishing the target from the background. Therefore, a number of follow-up studies based on this method have introduced improvements. In 2012, Henriques et al. [5] proposed the circulant structure of tracking-by-detection with kernel (CSK) algorithm based on MOSSE; in this approach, ridge regression was introduced as the loss function using the kernel technique. CSK optimized the training objective and derived the closed-form solution of the correlation filter. Furthermore, it greatly simplified the matrix multiplication operation in the Fourier domain by using the characteristics of the circulant matrix while retaining the speed advantage of MOSSE and achieving an improved tracking effect.

In 2014, Henriques et al. [6] improved the single channel filter of CSK to a multidimensional filter and proposed a new method called the kernel correlation filter (KCF) algorithm that updated the one-dimensional grayscale feature of CSK with the multidimensional histogram of oriented gradient (HOG) feature. Through experiments, this study fully verified that KCF substantially improved tracker performance.

In 2014, Danelljan et al. [7] proposed a new idea for scale-space filtering based on the correlation filtering algorithm that won the 2014 visual object tracking (VOT) competition. The algorithm is simple, with excellent performance and high portability. Compared with MOSSE, KCF, and other algorithms, the algorithm introduces two main contributions: a multifeature fusion mechanism and a relatively fast scale-space filtering optimization method.

More recently, in 2017, Danelljan et al. [8] proposed a new dimension-reduction tracking method, the efficient convolution operators (ECO), which simplifies feature extraction. Simplifying the feature set greatly reduced the calculations. Furthermore, the algorithm also stores historical target features in a manner that simplifies the feature set, which greatly improved tracking robustness.

Although tracking systems based on correlation filtering have made great progress, they have failed to make a breakthrough in efficient scale-space filtering [9] while maintaining tracking robustness. In the past three years, many researchers have investigated various tracking problems using neural networks [10], filter integration [11], filter channelization [12], deep learning [13], and other approaches. These newer studies provided many good ideas for our work.

In this paper, we present a tracking method that effectively estimates target scale by training a scale classifier. After determining the optimal target location, the target scale can be estimated independently. This method improves precision through an efficient scale-space search. The improvements of this article mainly include the following: (1) We propose a spatial filtering prediction technology based on time domain correlation. Using this technology substantially reduces the calculations needed for spatial filtering. (2) We also propose an adaptive overlapped filter strategy that avoids many unnecessary filter update calculations.

The paper is organized as follows. The second chapter mainly introduces the principle and skills of the discriminant correlation filter in the tracking algorithm. These principles and techniques will run through the following algorithm improvements. The third chapter introduces the proposed improvement method. In chapter four, a series of experiments compared with some existing traditional typical tracking algorithms are carried out to demonstrate the efficiency of the proposed method.

2. Learning Discriminant Correlation Filters

Current tracking algorithms are composed of five main parts: motion modeling, feature extraction, observation modeling, template or filter updating, and postprocessing [14]. Motion modeling is used to model the target motion. By predicting the target's position in the next frame, the corresponding target candidate search area can be obtained. Feature extraction uses a series of feature vectors to represent candidate image data while removing redundant features and retaining effective features. Observation modeling uses features extracted from candidate images to determine whether they represent the target or the background. Updating controls the strategy and frequency of model updating and balances model adjustment and tracking migration to maintain tracking accuracy while considering tracking robustness. When a tracking system includes multiple trackers, postprocessing integrates the results of each tracker into a final optimal tracking state.

The MOSSE tracker learns a discriminant correlation filter to locate a new frame's target position [4]. The method uses a series of image patches of the target's appearance (r_1, r_2,…, r_t) as training samples. These sample tags are associated with the filter output (out_1, out_2,…, out_t). The goal is to find a filter that maximizes the response to the target, that is, to satisfy $\text{out} = r \odot h$. The optimal correlation filter OCF_t for a time series t is obtained by the sum of the minimum mean square error:

$$\varepsilon = \sum_{i=1}^{t} \|\text{OCF}_t \odot f_i - \text{out}_i\|^2 = \frac{1}{MN} \sum_{i=1}^{t} \left\|\bar{H}_t F_i - G_i\right\|^2. \quad (1)$$

where the functions r_i, out_i, and h_t all have $M \times N$ dimensions. The symbol $\odot$ indicates a cyclic correlation. By $0 = (\partial/\partial H_t)\sum_{i=1}^{t}\|\bar{H}_t F_i - G_i\|^2$, Equation (1) can be minimized by the following filter model:

$$H_t = \frac{\sum_{i=1}^{t} \overline{G_i} F_i}{\sum_{i=1}^{t} \overline{F_i} F_i}. \quad (2)$$

The relevant output, out_i, is constructed by a Gaussian function whose peak value lies at the target position, r_i. In Equation (2), the numerator $\sum_{i=1}^{t} \bar{G}_i F_i$ and denominator $\sum_{i=1}^{t} \overline{F_i} F_i$ of H_t are updated separately by the weighted means of the new observations r_t.

Given a new image z, $y = \Psi^{-1}\{\bar{H}_t Z\}$ is used to calculate the correlation score y, where Ψ^{-1} denotes the inverse discrete Fourier transform (DFT) operation. The new target

position can be estimated by the maximum result of the correlation score y. The fast Fourier transform (FFT) is used to implement efficient training and searching.

On this basis, the ridge regression method is introduced into the least square problem of Formula (1) to form the biased estimation regression method. By giving up the unbiasedness of the least square, we can obtain an optimal filter fitting method with better tolerance to ill-conditioned data and more reliable calculations.

Let f represent a rectangular region of the target extracted from the feature image. The dimension of f is expressed as f^j. The goal is to find the optimal correlation filter ocf, which consists of one filter ocf^j for each feature dimension. The optimal correlation filter ocf can be obtained by minimizing the following loss function:

$$\varepsilon = \left\| \sum_{j=1}^{p} \text{ocf}^j \odot f^j - \text{out} \right\|^2 + \lambda \sum_{j=1}^{p} \left\| \text{ocf}^j \right\|^2, \tag{3}$$

where out is the relevant output of training sample f and the parameter $\lambda \geq 0$ controls the regularization. The solution to Equation (3) is

$$H^l = \frac{\bar{G} F^l}{\sum_{k=1}^{d} F^k F^k + \lambda}. \tag{4}$$

Adding the regularization term helps avoid overfitting and ill-conditioned solutions. Equation (4) still involves a matrix inversion process. Samples can be obtained by using the cyclic shift matrix properties to avoid inversion. Assume that $F = \Gamma \operatorname{diag}(\hat{f}) \Gamma^H$,

$$H = \Gamma \cdot \operatorname{diag}\left(\frac{\hat{f}}{\hat{f} \odot f \wedge^* + \lambda I} \right) \cdot \Gamma^H \cdot g. \tag{5}$$

Then, the convolution property of a cyclic matrix can be used to obtain the frequency domain display solution. Another advantage of the cyclic matrix method is its selection of positive and negative samples. The tracking algorithm trains the classifier by online learning. In each frame, appropriate positive samples and negative samples are selected for classifier training. Generally, the labels of negative samples are assigned a 0, while the labels of positive samples are assigned a 1. Theoretically, when more samples are collected, the tracker discrimination ability will be stronger. However, due to the time sensitivity of tracking, a modern tracker has to balance the number of samples with the amount of computation. Therefore, the common approach is to select only a small number of samples randomly from each frame. However, an insufficient number of samples adversely affect the tracker's judgment. When the circular matrix method is used, a sample can be used as a "base sample" to "copy" thousands of similar samples through circular displacement, but these samples are related to the base samples only after they are transferred to the Fourier domain, that is, these samples will not cause a calculation increase when they are calculated in the Fourier domain. Therefore, this method can obtain a nearly unlimited number of samples from an image without introducing large amounts of extra computation.

3. The Proposed Efficient Scale-Space Filtering

In this chapter, Section 3.1 briefly introduces the multiscale tracking method based on the feature pyramid; Section 3.2 gives the basic process of multiscale tracking based on the interactive iteration of position filter and scale filter. On this basis, it is pointed out that the main calculation cost of the algorithm will be greatly increased due to the introduction of a scale filter. In Section 3.3, a precise scale estimation method is introduced, and the search range of the scale level is greatly reduced by filtering and predicting the target scale on the time axis; in Section 3.4, in order to reduce the calculation process and tracking drift further, the improved algorithm changes the iterative calculation process of the position filter and scale filter and proposes an adaptive overlapping correlation filtering method based on deviation prediction on time scale.

3.1. Standard Scale-Space Tracking. An improved tracking method is based on learning a three-dimensional-scale spatial correlation filter. The scale of this filter is fixed at $X \times Y \times Z$, where Z is the number of scales. To update the filter, the feature pyramid of the rectangular region around the target is calculated. To obtain the target position in the new frame, the $X \times Y \times Z$ rectangular cuboid is extracted from the feature pyramid as described above. In theory, the more samples that are collected, the stronger the discriminative ability of the trained tracker is. However, due to the time sensitivity of tracking, the tracker must balance the number of samples with the computational complexity.

3.2. Discriminative Scale-Space Tracking. Two correlation filters (a position filter and a scale filter) are used for target location and scale evaluation. The two filters are relatively independent; therefore, different feature types and feature calculation methods can be selected during their training and testing. Training sample f, which is used to update the scale filter, is obtained by extracting features of different image sizes around an object at the center. Similar to the spatial filter and the scale filter defined in Section 3.1, we can extract a target image at the center position and a size of s, where a represents the scale factor. The iterative process of the scale filter and spatial filter undoubtedly improves the efficiency of the scale-space search. Because the scale parameter is an exponential function, the scale size does not increase linearly: the larger the scale parameter is, the larger the search step is. In contrast, the smaller the scale parameter is, the finer the scale-space searches are. That is, coarse detection is conducted at a larger scale, while fine detection is conducted at a smaller scale.

3.3. Scale-Space Search Based on Time Association. The introduction of scale estimation into the tracker leads to a greater computational cost. Ideally, an accurate scale estimation method should be both robust and efficient simultaneously.

Table 1: Analysis of traditional tracking in the UAV123 database without scale estimation (showing some videos with typical tracking problems).

Video name	Tracking failure location	Failure reason description	Video name	Tracking failure location	Failure reason description
Car1	1/3	Scale+background change	Person12	3/5	Shelter
Car3	4/5	Background accumulation	Person14	1/10	Shelter
Car6	1/3	Out of sight	Person16	1/5	Shelter
Car9	2/5	Occlusion+scale change	Person21	1/100	Small target
Car11	1/100	Small target	Person1_s	1/3	Similar target
Car15	1/100	Small target	Truck1	1/2	Illumination change and deformation
Car16	1/5	Rapid deformation+scale change	Truck2	1/3	Shelter
Person4	1/3	Scale change, severe deformation	Wakeboard3	1/3	Scale+deformation+rapid movement
Person6	9/10	Deformation and scale accumulation	Wakeboard5	1/3	Rapid movement
Person9	1/5	Out of sight	Wakeboard8	1/3	Scale change+background learning

In a visual tracking scene, the scale difference between two frames is usually smaller than that of the positional difference because the relative distance between the target and the camera in the tracking process generally does not change dramatically over short interframe periods. For the target-tracking process in a monitoring scene, in which the camera is typically fixed and the target moves, the target scale changes are mainly caused by the advance or retreat of the target. In contrast, during the dynamic tracking process when the camera is in motion but the target remains relatively still, scale changes to the target are mainly caused by changes in the optical parameters of the camera. However, the scale changes produced by these processes have their own specific laws. We use the Kalman filter method to filter and predict the target scale on the time axis, which greatly reduces the search range at the scale level. The adaptive filter is used to control the size of the search interval. When the residual is reduced, the search range is narrowed; conversely, when the residual is larger, the search range is enlarged.

3.4. Adaptive Cross-Correlation Filtering. When scale filtering is added to the traditional spatial filtering target-tracking process, the entire filtering calculation process is multiplied, and the real-time advantage of correlation filtering is significantly reduced. To reduce unnecessary calculations and improve the efficiency of the complete filtering process, we propose an adaptive overlapping correlation filtering method. In this filtering process, the traditional alternating calculations of the position and scale filters in each frame are changed to calculate either only the position filter or only the scale filter in a single frame. The specific filtering process depends on the deviation predictions of the position filter and the scale filter in the time scale. Generally, when the target's position changes rapidly, the scale change is usually small. In contrast, when the scale changes drastically, the position is usually undisturbed. Finally, when both the position and the scale changes are large, the position filter and scale filter should be memorized for a brief period instead of being refreshed quickly to reduce the tracking drift.

4. Experiments and Analysis

In this study, we compare our novel and fast scale estimation-based tracking method with existing traditional typical tracking algorithms.

4.1. Analysis of the Test Dataset. The commonly used datasets for target tracking include the OTB and VOT series; however, larger datasets are also currently available: LASOT and TrackingNet. In addition, there is a dataset called UAV123, which is a special scene dataset, all of whose images were acquired by unmanned aerial vehicles (UAVs). These images are characterized by clean backgrounds and numerous perspective changes. The dataset has a total size of approximately 13.5 Gb.

First, we tested the common correlation filtering algorithms on the UAV123 database without scale estimation. Analysis of traditional tracking in the UAV123 database without scale estimation is shown in Table 1. As seen from Table 1, the reasons for target-tracking failures in the image sequence mainly involve background error accumulation, scale changes, rapid scale changes, occlusion, small targets, similar targets, fast movement, illumination changes, and the target moving out of the image area.

By further investigating the above reasons for errors, we find the following:

(1) A small target is usually related to initialization failure; a small target has few pixels and little feature information; thus, tracking fails quickly either during initialization or close to the start of tracking

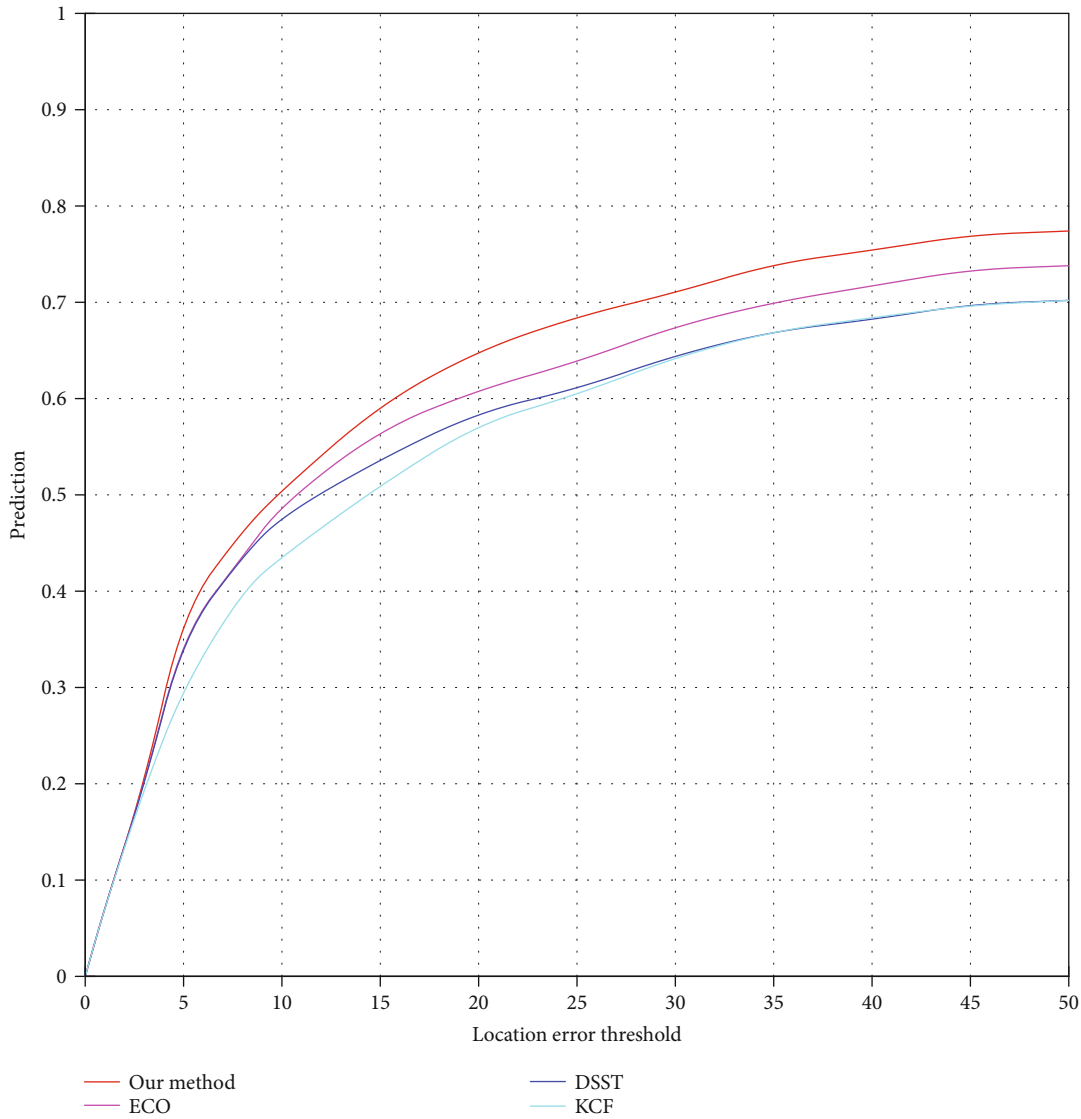


Figure 1: Accuracy curves of the compared algorithms on the UAV123 dataset.

(2) For occlusions and situations in which the target leaves the image area, tracking inevitably fails. This condition needs to be solved by adding a target redetection module

(3) Simultaneous rapid movement and target deformation will also lead to target-tracking failure. This is because the size of the search area is limited. When rapid movement occurs, the center of the target is displaced by a large amount, which may cause the target to cross the search boundary. When rapid deformation occurs, the appearance information of the target changes substantially, allowing the filter insufficient learning time; this leads to target-tracking failure

(4) Correlation filtering is able to resist changes in illumination, but when the target illumination suddenly changes dramatically, such as when the target enters a shadowed area from sunlight, tracking may fail

(5) Regarding the influence of similar targets, when two similar targets are close together or cross paths, the filter may fail to track the correct target

(6) Background error accumulation and scale change are also issues. Because the selected target area includes some background information, when the target appearance or scale changes, the filter will continuously learn more background information. Consequently, over time, these errors accumulate, eventually resulting in tracking failure

4.2. Experimental Configuration. The performance of the proposed algorithm is verified quantitatively by standard experimental parameters [15]. The experimental results are

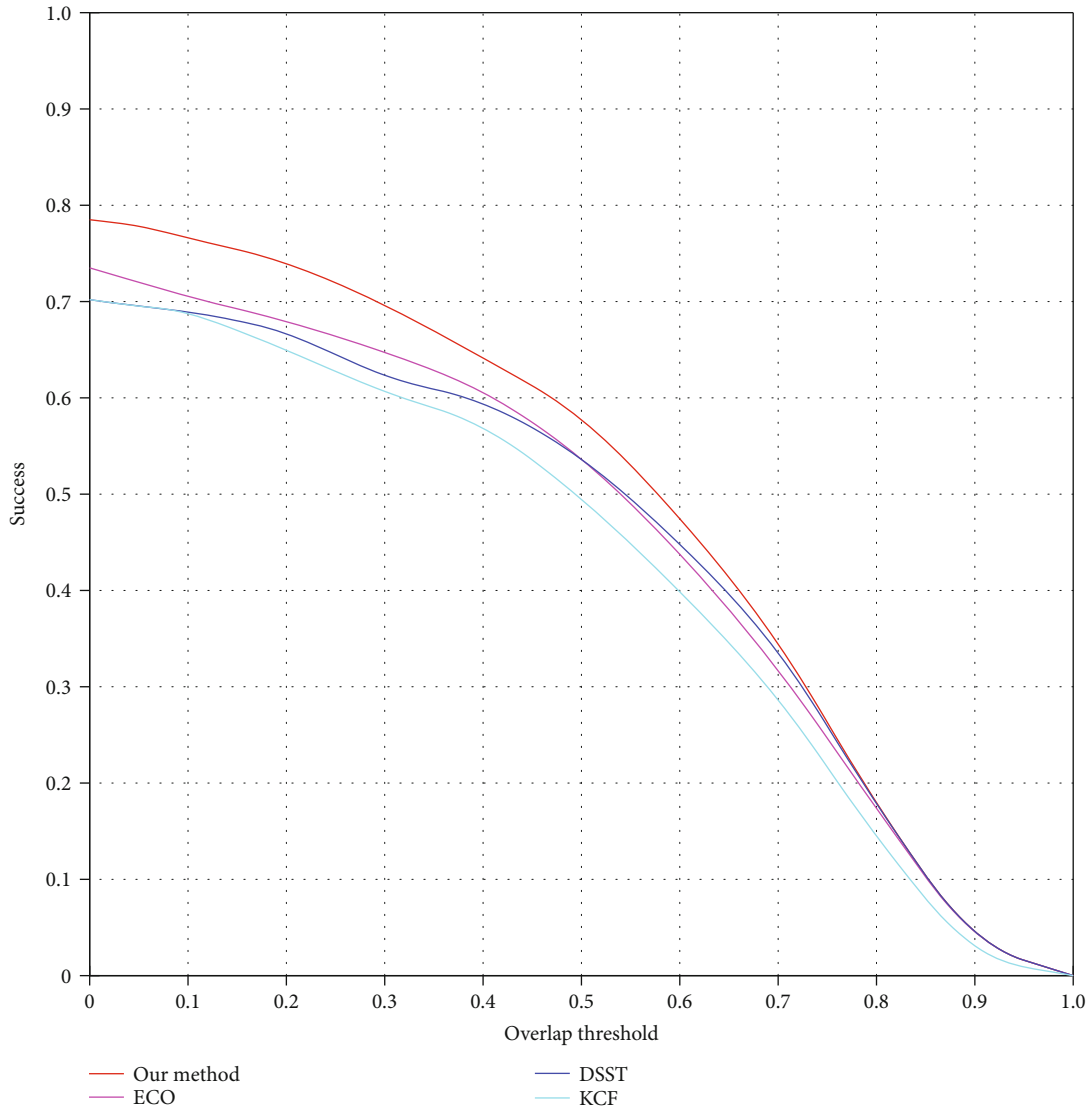


Figure 2: Success rate curves of the compared algorithms on the UAV123 dataset.

represented by the metrics of distance accuracy (DP), central position error (CLE), and overlap accuracy (OP). Additionally, the tracker's speed is represented by the median frame rate (FPS) in the video frames. We also report the results in terms of accuracy and through success rate graphs [10].

The DP is the ratio of the number of test video frames whose central position error is less than a threshold number of pixels to the total number of test video frames. This index represents the overall stability of the visual tracking process.

CLE represents the Euclidean distance between the tracking algorithm output target center and the calibrated target center. This index reflects the degree of coincidence between the tracking center and the actual target center, and it is one of the descriptive indexes of tracking accuracy.

In the tracking algorithm, the OP of each video frame is equal to the ratio of the intersection between the output frame area and the calibration frame area to the area of the union between the output frame area and the calibration frame area. Generally, an overlap ratio greater than 0.5 represents tracking success for the current frame; otherwise, it is considered a tracking failure. This index indicates the overall correctness of the visual tracking process.

FPS reflects the rapidity of the tracking algorithm.

4.3. Comparison with Traditional Tracking Algorithms. We performed this experiment on the UAV123 dataset. This dataset consists of many video sequences depicting a variety of scenes, such as biking, boating, driving a car, groups of people, trucks, individuals, and wake boarding, all of which were acquired using a UAV. The dataset includes changes in target scale, fast-moving targets, small targets, target shape changes, and target occlusions.

The comparison algorithms include our proposed algorithm, the ECO algorithm, the visual tracking via discriminative sparse similarity map (DSST) [16] algorithm, and the KCF algorithm.

The parameters are set as follows: padding: 4; HOG features: cell_size = 6, hog_orientations = 9; compression

Table 2: Comparison of algorithm accuracy and speed in the UAV123 dataset.

Methods	DSST	KCF	ECO	Our method
P20	0.582	0.569	0.609	0.649
AUC	0.579	0.565	0.605	0.638
FPS	28	42	24	39

dimension: 10; CN features: cell_size = 4; compression dimension: 3; CG iterations: 5; ideal Gaussian output Sigma: 1/16; learning rate: 0.009; and sample space size: 30.

Figures 1 and 2, respectively, show the tracking accuracy curves and success rate curves of the four algorithms (including our new algorithm) on the UAV123 dataset. As the figures show, the tracking accuracy and success rate of the proposed algorithm are obviously better than those of the other algorithms. This result occurs primarily because the correlation filtering in the new algorithm uses the characteristics of the target area to "match" in the next frame, and these good characteristics make the target and the complex background easier to discriminate. In addition, the new algorithm has high discrimination and robustness capabilities for both multidimensional information features and spatial scale features. Table 2 also shows that in the tracking accuracy P20 data, the new algorithm achieves 0.649, which is significantly higher than the other algorithms. Regarding the success rate shown by the area under the curve (AUC) data, the new algorithm achieves 0.638, which is also significantly higher than the other algorithms. For FPS, the highest score is 42, by KCF, but the new algorithm achieves a score 39, which is not considerably different from the KCF score, showing that the new algorithm does not significantly increase the amount of computation after adding the efficient scale filtering using the strategy described above.

5. Conclusion

In this paper, we present a novel scale estimation algorithm for visual tracking. This article uses a visual tracking algorithm based on a robust scale estimation process, which uses a discriminant correlation filter based on a time-dependent scale-space filter and an adaptive cross-correlation filter. Compared with the traditional filter approach, the proposed tracker provides a better overall performance and improves the computational efficiency. Moreover, because the scale estimation process presented in this paper is independent, it can easily be introduced into any tracking algorithm. Our algorithms show good stability and robustness. In the future work, we plan to study how to address tracking tasks in more complex environments and improve the capability of visual tracking in quickly changing scenes.

Data Availability

The data used to support the findings of this study are available from this website: https://cemse.kaust.edu.sa/ivul/benchmark-and-simulator-uav-tracking-dataset(UAV123).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is financially supported by the Cultivation Program for Youth Backbone Teachers in Colleges and Universities of Henan Province (Grant No. 2019GGJS184) and the Key Technologies R&D Program of Henan Province (Grant No. 182102310752).

References

[1] Y. Sakai, H. Lu, J.-K. Tan, and H. Kim, "Recognition of surrounding environment from electric wheelchair videos based on modified YOLOv2," *Future Generation Computer Systems*, vol. 92, pp. 157–161, 2019.

[2] H. Lu, Y. Li, T. Uemura, H. Kim, and S. Serikawa, "Low illumination underwater light field images reconstruction using deep convolutional neural networks," *Future Generation Computer Systems*, vol. 82, pp. 142–148, 2018.

[3] Y. Zhang, R. Gravina, H. Lu, M. Villari, and G. Fortino, "Pea: parallel electrocardiogram-based authentication for smart healthcare systems," *Journal of Network and Computer Applications*, vol. 117, pp. 10–16, 2018.

[4] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2544–2550, San Francisco, CA, 2010, IEEE.

[5] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *European conference on computer vision*, pp. 702–715, Florence, Italy, 2012, Springer.

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[7] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference*, Nottingham, 2014BMVA Press.

[8] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: efficient convolution operators for tracking," in *IEEE Conference on Computer Vision & Pattern Recognition*, Los Alamitos, CA, USA, 2017.

[9] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561–1575, 2017.

[10] Z. Zhu, G. Huang, W. Zou, D. Du, and C. Huang, "Uct: learning unified convolutional networks for real-time visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1973–1982, Venice, 2017.

[11] B. Uzkent and Y. Seo, "EnKCF: Ensemble of kernelized correlation filters for high-speed object tracking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1133–1141, Lake Tahoe, NV, USA, 2018, IEEE.

[12] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition, pp. 6309–6318, Honolulu, HI, USA, 2017.

[13] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2805–2813, Honolulu, HI, USA, 2017.

[14] R. Wenbi, R. Chunyang, and X. Qiang, "Correlation filtering target tracking based on color and part spatial relation constraints," in *Proceedings of the 2nd International Conference on Intelligent Information Processing -IIP'17*, Bangkok, Thailand, 2017.

[15] K. Zhang, Q. Liu, Y. Wu, and M.-H. Yang, "Robust visual tracking via convolutional networks without training," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1–1792, 2016.

[16] B. Zhuang, H. Lu, Z. Xiao, and D. Wang, "Visual tracking via discriminative sparse similarity map," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1872–1881, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8894760, 13 pages
https://doi.org/10.1155/2020/8894760

Research Article

A Multichannel Biomedical Named Entity Recognition Model Based on Multitask Learning and Contextualized Word Representations

Hao Wei, Mingyuan Gao, Ai Zhou, Fei Chen, Wen Qu, Yijia Zhang, and Mingyu Lu

School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

Correspondence should be addressed to Mingyu Lu; lumingyu@dlmu.edu.cn

Received 24 May 2020; Revised 15 June 2020; Accepted 30 June 2020; Published 10 August 2020

Academic Editor: Yin Zhang

As the biomedical literature increases exponentially, biomedical named entity recognition (BNER) has become an important task in biomedical information extraction. In the previous studies based on deep learning, pretrained word embedding becomes an indispensable part of the neural network models, effectively improving their performance. However, the biomedical literature typically contains numerous polysemous and ambiguous words. Using fixed pretrained word representations is not appropriate. Therefore, this paper adopts the pretrained embeddings from language models (ELMo) to generate dynamic word embeddings according to context. In addition, in order to avoid the problem of insufficient training data in specific fields and introduce richer input representations, we propose a multitask learning multichannel bidirectional gated recurrent unit (BiGRU) model. Multiple feature representations (e.g., word-level, contextualized word-level, character-level) are, respectively, or collectively fed into the different channels. Manual participation and feature engineering can be avoided through automatic capturing features in BiGRU. In merge layer, multiple methods are designed to integrate the outputs of multichannel BiGRU. We combine BiGRU with the conditional random field (CRF) to address labels' dependence in sequence labeling. Moreover, we introduce the auxiliary corpora with same entity types for the main corpora to be evaluated in multitask learning framework, then train our model on these separate corpora and share parameters with each other. Our model obtains promising results on the JNLPBA and NCBI-disease corpora, with F1-scores of 76.0% and 88.7%, respectively. The latter achieves the best performance among reported existing feature-based models.

1. Introduction

Named entity recognition (NER) aims to identify and extract specific entities (persons, places, organizations, and so on) from massive unstructured text data, which becomes a primary task for information extraction, text analysis, text mining, etc. Similarly, how to effectively extract and obtain valuable information has become a serious challenge for researchers in the biomedical field. Biomedical named entity recognition (BNER) is an indispensable step for this above challenge. The biomedical entities consist of genes, proteins, diseases, drugs, chemicals, and so on.

In the past, conventional machine learning methods were widely used for NER, such as support vector machine (SVM), conditional random field (CRF), and maximum entropy model (MEM). Finkel et al. [1] combined distant resources and additional features to identify the biomedical entities. Tsuruoka et al. [2] employed MEM to develop a BNER system named GENIA Tagger. ABNER [3] was a biomedical entities extraction system based on CRF. Chang et al. [4] adopted the biomedical word embeddings as external features to improve the performance of CRF significantly. Liao et al. [5] adopted the Skip-Chain CRF model to recognize entities, which effectively captured the features of the distant context. Tang et al. [6] used a CRF model with three different types of word representations to identify biological entities. According to the above studies, CRF had became the mainstream model in BNER [7]. Nevertheless, feature engineering is an essential element of the conventional machine learning methods. They must manually

design complex templates that require not only domain knowledge but also time-consuming.

Driven by artificial intelligence and pattern recognition, some labor-saving and advanced technologies have been developed in natural language processing, computer vision, and other emerging fields [8–17]. For example, deep learning can obviously address the expensive cost of feature engineering. The widely employed neural networks include convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and gated recurrent unit networks (GRUs). Yao et al. [18] first built a multilayer neural network to obtain the biomedical word embeddings on large-scale corpora. To extract disease and chemical entities, Zhao et al. [19] constructed a CNN model. In this work, BNER was seen as text classification, and a multilabel mechanism was designed to obtain contiguous labels. Zhu et al. [20] adopted a CNN structure in BNER with n-gram local character and word embeddings. The GRAM-CNN obtained the best performance (F1-score: 87.3) among the single-task models on the NCBI-disease corpus. Li et al. [21] made connections between the twin word embeddings and sentence vectors. Furthermore, they adopted the bidirectional LSTM (BiLSTM) to identify biomedical entities and significantly improved the performance. Limsopatham et al. [22] proposed an end-to-end model based on BiLSTM and orthographic features. It was designed to improve the extraction of complex biomedical terms. SBLC was developed by Xu et al. [23] based on word embeddings and BiLSTM-CRF structure. Dang et al. [24] also proposed the BNER model based on the BiLSTM-CRF structure and adopted various fine-tuned linguistic embeddings. The model showed high performance on multiple corpora. Lyu et al. [25] adopted the BiLSTM-RNN model and combined the biomedical word embeddings with character embeddings to recognize entities. In addition, some studies based on multitask learning and transfer learning were widely used in BNER and had achieved competitive performance. Wang et al. [26] jointly trained different types of entities in multiple data sets and shared both word and character representations among relevant entities. The multitask model achieved promising performance on 15 biomedical corpora. Yoon et al. [27] proposed a multitask framework termed CollaboNet. It connected multiple submodels trained on different corpora. The large performance gains come from taking turns training the target and collaborator submodels. Sachan et al. [28] designed a pretrained BiLSTM model. They first trained a language model of the same structure on the unlabeled corpora and then updated the initialization parameters of the BNER model based on transfer learning. It does not only substantially improved the performance but also alleviated the lack of high-quality labeled training data.

From the above studies, word embeddings can be seen to have become indispensable representations. They can effectively represent the semantic features of the original text sequences. But biomedical entities' naming rules are vague. There are many polysemous and ambiguous words in the biomedical literature. For example, in "This cohort underwent follow-up for cancer incidence through the Finnish cancer registry to the end of 1995.", the first "cancer" means disease and the second is an institution. In addition, it is difficult to address the lack of sufficient training samples in specific fields. These issues also result that the biomedical entities are more complex to recognize than the general field. Because the traditional fixed word embeddings cannot accurately represent polysemous and ambiguous words in the biomedical literature, the language models pretrained on a large number of unlabeled open corpora have drawn more and more attention. The contextualized word embeddings generated by them can optimize the feature representations of the polysemous and ambiguous words. In the general field, Peters et al. [29] designed a feature-based language model named ELMo, which consists of a bidirectional LSTM. This pretrained language model achieves state-of-the-art performance in multiple downstream tasks.

We aim to optimize the representations of polysemic words and ambiguous words in biomedical sequences and make the model fully capture richer features. This paper proposes a multitask learning multichannel BiGRU-CRF model with feature-based contextualized word representations. The main contributions of this paper are as follows.

1) We propose a multichannel BiGRU-CRF model. Three kinds of feature representations based on the biomedical pretrained dictionary, ELMo, and CNN are generated, including word-level, contextualized word-level, and character-level representations. These representations are separated or combined as inputs simultaneously, and each set of inputs is fed into a BiGRU-CRF model as a single channel. In merge layer, multiple methods are designed to integrate the outputs of multichannel BiGRU.

2) In order to address the lack of sufficient training data in specific fields, we adopt multitask learning strategy, employing auxiliary corpora to provide richer training samples and relevant information for the main corpora to be evaluated.

3) The multitask learning multichannel BiGRU-CRF model clearly strengthens the capability of recognizing entities without any artificial participation. It obtains the competitive results on the JNLPBA and NCBI-disease corpora.

The rest of this paper is divided into the following four sections. Section 2 describes the methods. Section 3 shows the experimental settings. Section 4 reports the evaluative results in a detailed manner. Section 5 provides the conclusion.

2. Methods

Figure 1 shows the multitask learning multichannel BiGRU-CRF framework. The framework is divided into five parts: input layer, embedding layer, BiGRU, merge layer, and CRF layer, where the input layer represents the original sentence in corpora. First, the three feature representations are obtained through biomedical pretrained dictionary, CNN, and ELMo language model, respectively. Then, the multichannel BiGRU is used to capture features. $\vec{h}_{0-6}$ denotes the forward single-channel GRU, and $\overleftarrow{h}_{0-6}$ denotes the backward single-channel GRU, respectively. Next, we integrate the output of each channel in the merge layer. Finally, the labels are parsed by CRF. This section describes the remaining four parts in detail.

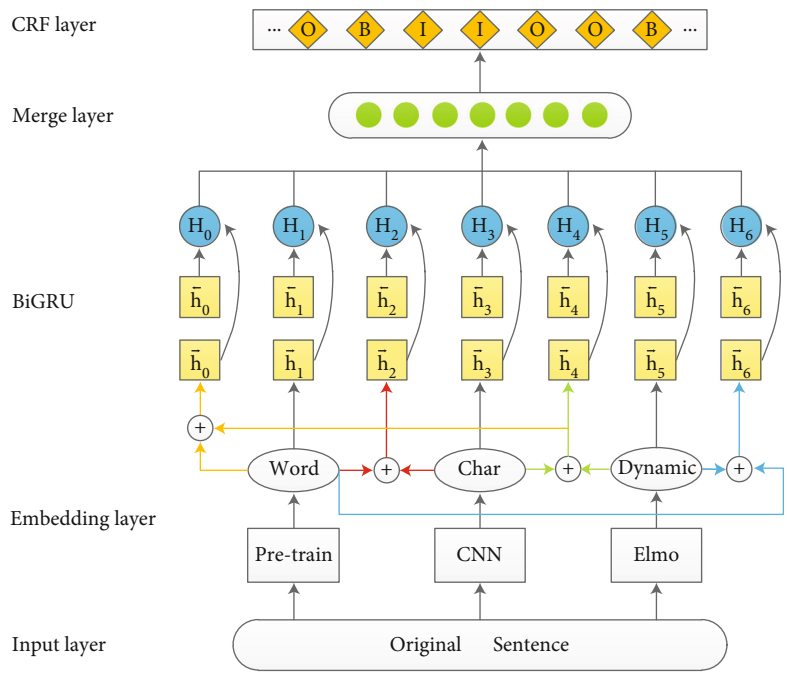


Figure 1: Multichannel BiGRU-CRF architecture. It consists of 5 parts: input layer, embedding layer, BiGRU, merge layer, and CRF layer. The red, green, and blue lines, respectively, represent the channels after the concatenate operation of two representations, and the yellow line represents the channel after the concatenate operation of all three representations. $H0$-6, respectively, denotes the bidirectional single-channel GRU. Each channel is independent, which avoids information redundancy.

2.1. Embedding Layer. To ensure the maximum coverage of the input information, the pretrained word embeddings, contextualized word embeddings, and character embeddings are used for the input layer for feature representations.

2.1.1. Pretrained Word Embedding. We represent the text sequence with word embeddings. They map words to dense vectors according to semantic relevance. The word embedding method addresses the lack of curse of dimensionality compared with the conventional one-hot method. With the development of natural language processing, word embeddings have become the most important input feature representations. The widely adopted word embedding computing tools include Word2Vec [30] and GloVe [31].

Previous biomedical studies have provided related open source word embeddings pretrained on large-scale unlabeled corpora. We initialize the word embeddings by a "look up" operation. Inspired by Quan et al. [32], this paper adopts the word embeddings pretrained on *PMC* and *PubMed* biomedical corpora.

2.1.2. Contextualized Word Embedding. This paper directly transfers the pretrained ELMo language model proposed by Peters et al. [29] to obtain the contextualized word embeddings. The main motivation is that the contextualized word representations should be able to contain rich syntactic and semantic information. The conventional word embeddings (e.g., word2vec) are context-independent, and ELMo can generate dynamic word embeddings based on context. We adopt the 2-layer ELMo to obtain the contextualized word representations as part of the multichannel BiGRU-CRF model's input, which is shown in Figure 2. ELMo consists of a bidirectional LSTM language model. The objective function is to compute the maximum likelihood of the two submodels. For $k-th$ word, a set of contextualized word representations can be computed by ELMo as follows:

$$\begin{aligned} ELMo_k &= \sum_{j=0}^{L} w h_{k,j}^{LM} \\ R_k &= \left\{ x_k^{LM}, h_{k,j}^{LM\vec{}}, h_{k,j}^{LM\overleftarrow{}} \right\}, j = \{1,\cdots,L\} \\ R_k &= \left\{ h_{k,j}^{LM} \right\}, j = \{0,\cdots,L\} \end{aligned} \tag{1}$$

where x_k^{LM} denotes the original embeddings layer. $h_{k,j}^{LM\vec{}}$ and $h_{k,j}^{LM\overleftarrow{}}$ denote the forward and backward LSTM layer, respectively. w denotes the softmax-normalized weights, and L denotes the number of layers. ELMo generates word representations based on the above formula, which is summing each hidden state of the bidirectional language model. They can be directly concatenated with other feature inputs. The contextualized word embeddings not only reflect the complex semantics and grammar features but also accurately adapt to different contexts.

2.1.3. Character Embedding. Character representations refer to morphological information by capturing it from all characters that make up a word. Combining them with other

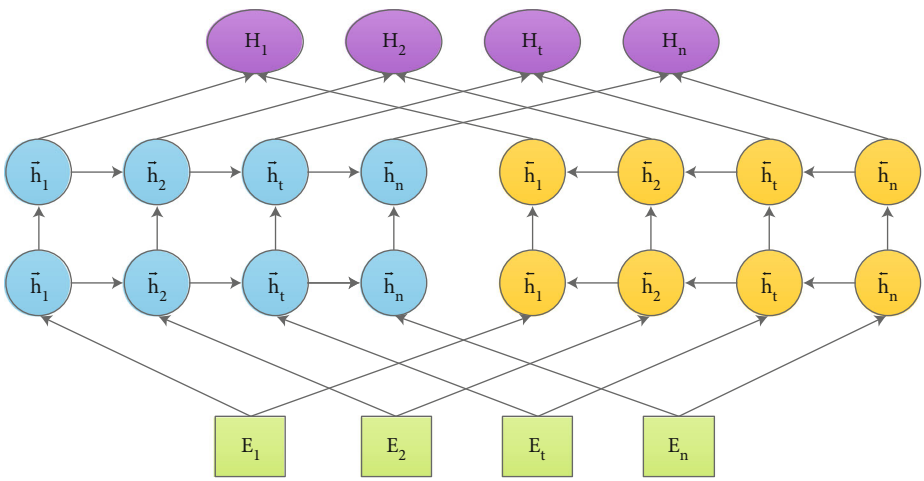


Figure 2: The framework of ELMo.

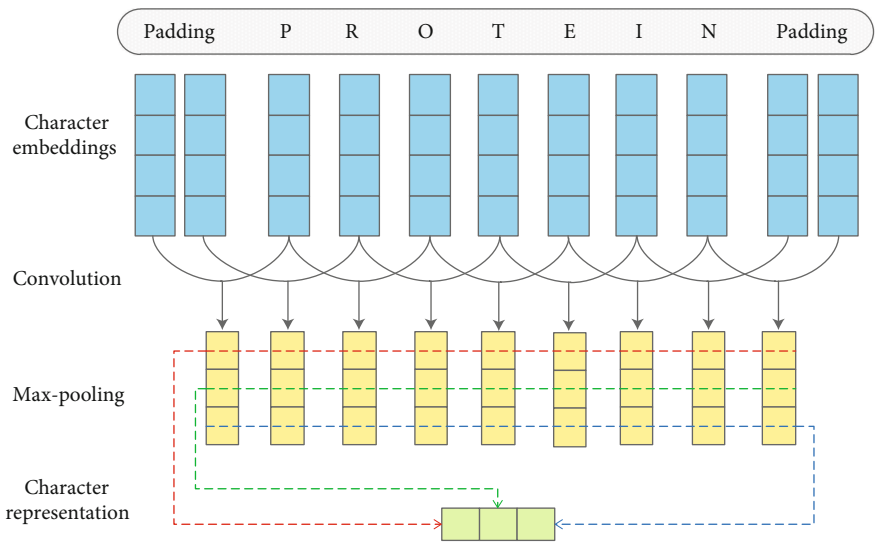


Figure 3: The CNN model framework.

feature representations can better describe the morphological features of a word [33, 34]. Previous studies have shown the effectiveness of character representations in NER. This paper adopts CNN to compute the character vectors of words in biomedical sequences. The structure of CNN is shown in Figure 3, including the original character embeddings by random initialization, convolutional layer, and pooling layer. First, the words' embeddings matrix consists of each character embeddings. A padding operation for words of different lengths is performed. Then, the local features of the initialized character embeddings matrix are captured by a convolution operation. Finally, the character representations are obtained by performing a max-pooling operation.

2.2. Multichannel BiGRU. Recently, to solve the gradient explosion or gradient disappearance, a variety of improved models based on RNN have been proposed, such as LSTM [35] and GRU [36–38]. They capture distant information and address the gradient disappearance or gradient explosion by designing the memory units and gate mechanisms. Therefore, the above improved models have become the major option for sequence labeling such as BNER. The difference between LSTM and GRU is the structure of gate mechanisms. GRU maintains the performance of LSTM while making the gate structures simpler [39, 40]. Because we need to train multiple identical networks at the same time, this paper adopts GRU with lower computational complexity. Figure 4 shows the GRU units. The relevant formulas are as follows.

$$\begin{aligned} z_t &= \sigma(W_z[h_{t-1}, x_t]) \\ r_t &= \sigma(W_r[h_{t-1}, x_t]) \\ h_t^{\sim} &= \tanh\left(W[r_t * h_{t-1}, x_t]\right) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * h_t^{\sim} \end{aligned} \tag{2}$$

where σ denotes the *sigmod* function. z_t and r_t denote the update and reset gate. x_t denotes the feature vectors. W denotes the weights of the gate mechanism. $h_t^{\sim}$ denotes the

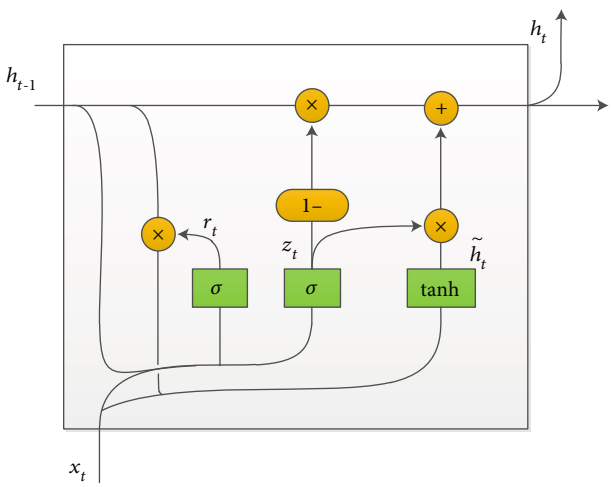


Figure 4: The unit of GRU.

current state. tanh denotes the hyperbolic tangent function. h_t denotes the final output.

However, GRU only considers the forward information of texts and ignores the backward information, which also contains important features. The bidirectional GRU is employed in our model because of this issue. The BiGRU model captures different bidirectional feature representations in each sequence. Then, it obtains the complete representations by connecting them. BiGRU can capture the bidirectional representations and hidden features. In our model, we propose a multichannel BiGRU to obtain the richer representations. The multichannel mechanism aims to feed different kinds of input representations into corresponding multiple independent and same network structures. Each channel uses a separate BiGRU to capture features, which does not cause interference between the channels and can extract information more adequately. A total of 7 channels are designed to capture features of different representations, as follows.

(1) 1st channel: pretrained word embeddings ⊕ contextualized word embeddings ⊕ character embeddings

(2) 2nd channel: pretrained word embeddings

(3) 3rd channel: pretrained word embeddings ⊕ character embeddings

(4) 4th channel: character embeddings

(5) 5th channel: contextualized word embeddings ⊕ character embeddings

(6) 6th channel: contextualized word embeddings

(7) 7th channel: pretrained word embeddings ⊕ character embeddings

where ⊕ denotes the concatenate operation.

2.3. Merge Layer. The purpose of using the merge layer is to integrate the outputs of multiple channels from BiGRU. A good merge scheme can effectively integrate the potential valuable information in multichannel BiGRU. As shown in Figure 1, the multichannel BiGRU is adopted to capture features from different representations. Let H' denotes the multichannel BiGRU's output. For a given text sequence $S = \{s_1, s_2, \cdots, s_m\}$, m denotes the length of the sequence, and u denotes the number of BiGRU units. We design four merge methods: addition, connection, unit-level attention, and channel-level attention.

1) Addition. This method additively integrates the output of each channel, and each single BiGRU does not interfere with others when capturing features. It can be obtained as follows:

$$\begin{aligned} H_i &= \left[\overrightarrow{h_i} \oplus \overleftarrow{h_i}\right] \\ H' &= H_w + H_e + H_c + H_{\mathrm{we}} + H_{wc} + H_{ec} + H_{wec} \end{aligned} \tag{3}$$

where + denotes element-wise addition, $H' \in m \times u$. H_i denotes the single-channel BiGRU's output, w, e, and c, respectively, denote the pretrained word embeddings, the contextualized word embeddings from ELMo, and the character embeddings from CNN.

2) Connection. This method directly performs the concatenate operation on the single-channel BiGRU's output. It can be obtained as follows:

$$\begin{aligned} H_i &= \left[\overrightarrow{h_i} \oplus \overleftarrow{h_i}\right] \\ H' &= H_w \oplus H_e \oplus H_c \oplus H_{\mathrm{we}} \oplus H_{wc} \oplus H_{ec} \oplus H_{wec} \end{aligned} \tag{4}$$

where ⊕ denotes the concatenate operation, $H' \in m \times 7u$. w, e, and c, respectively, denote the 3 different embeddings.

3) Unit-level attention. This method adopts the multi-head self-attention mechanism to redistribute the weights of units in BiGRU. It can be obtained as follows:

$$\begin{aligned} H_i &= \left[\overrightarrow{h_i} \oplus \overleftarrow{h_i}\right] \\ \alpha &= Softmax\left(\frac{QK^T}{\sqrt{u}}\right) \\ head_i &= \sum_m \alpha V \\ MH(Q, K, V)_i &= (head_1 \oplus \cdots \oplus head_H) \\ H' &= \sum_{i=1}^{n} MH_i \end{aligned} \tag{5}$$

where ⊕ denotes the concatenate operation. H_i denotes the single-channel BiGRU's output, $Q, K, V \in m \times (u/H)$, $MH_i \in m \times u$, $H' \in m \times u$.

4) Channel-level attention. This method first connects the feature representations of all channels, then computes the weights of each channel and finally integrates them. It

can be obtained as follows:

$$\begin{aligned} H_i &= \left[\vec{h_i} \oplus \overleftarrow{h_i}\right] \\ H &= H_w \oplus H_e \oplus H_c \oplus H_{\mathrm{we}} \oplus H_{wc} \oplus H_{ec} \oplus H_{wec} \\ \alpha_i &= \frac{\exp\left(e_i\right)}{\sum_{k=1}^{n} \exp(e_k)} \\ e_i &= \tanh\left(W^T H + b\right) \\ H' &= H \otimes \alpha_i \end{aligned} \tag{6}$$

where $\otimes$ denotes matrix multiplication, $H \in m \times u \times 7$, $H' \in m \times u$.

2.4. CRF Layer. After the representations information is output by BiGRU, the conventional decision function computes the prediction labels Y. However, the output sequence labels have strong dependence in BNER. For example, in the *BIO* labeling scheme, the previous label of "B-disease" cannot be "I-disease". The conventional decision function is insufficient to address the above issue effectively.

In our model, CRF [41] is employed after the merge layer; hence, the dependence between the output labels can be effectively considered. For sentence $X = \{x_1, x_2, \cdots, x_m\}$, it is input into BiGRU. P denotes the probability which is output from merge lager, $P \in m \times n$. m denotes the sequences, and n denotes the labels. p_{ij} denotes the $j-th$ label probability of the $i-th$ token. Y denotes the prediction labels, where $Y = \{y_1, y_2, \cdots, y_m\}$. Its probability can be obtained as:

$$P(X, Y) = \sum_{i=0}^{m} F_{y_i, y_{i+1}} + \sum_{i=1}^{m} P_{i, y_i} \tag{7}$$

where F denotes the transfer matrix. $F_{y_i, y_{i+1}}$ denotes the transition probability from y_i to y_{i+1}. The probability of all prediction labels Y by decision function can be computed as follows:

$$P(Y \mid X) = \frac{exp^{P(X,Y)}}{\sum_{Y^{\sim} \in Y_X} exp^{P\left(X, Y^{\sim}\right)}} \tag{8}$$

$Y^{\sim}$ denotes the truth labels.

The likelihood function is:

$$\log\left(P(Y \mid X)\right) = P(X, Y) - \log\left(\sum_{Y^{\sim} \in Y_X} exp^{P\left(X, Y^{\sim}\right)}\right) \tag{9}$$

Y_X denotes all legal label sequences. The final prediction label sequence with the maximum probability can be gained as follows:

$$Y^* = argmax_{Y^{\sim} \in Y_X} P(X, Y^{\sim}) \tag{10}$$

2.5. Multitask Learning. In order to provide more training data and value information for our model, we adopt the multitask learning strategy. The basic idea of multitask learning is to learn multiple tasks at the same time and use related information between tasks to improve model performance. The neural network-based multitask learning method mainly adopts a parameter sharing learning mode to learn a shared representation for multiple tasks. In this paper, we introduce two auxiliary corpora with the same entity types for the main corpora to be evaluated, then train the multichannel BiGRU-CRF model on these separate corpora and share parameters with each other.

Given a set of training corpus n, $n \in \{1, \cdots, n\}$. X_i and Y_i represent the samples and corresponding prediction labels in each corpus, respectively. The loss function L of the model based on multitask learning is as follows:

$$\begin{aligned} L &= \sum_{i=1}^{n} \alpha_i L_i \\ &= \sum_{i=1}^{n} \alpha_i \log\left(P(Y_i \mid X_i)\right) \\ &= \sum_{i=1}^{n} \alpha_i \left(P(X_i, Y_i) - \log\left(\sum_{Y_i^{\sim} \in Y_X} exp^{P\left(X_i, Y_i^{\sim}\right)}\right)\right) \end{aligned} \tag{11}$$

where α_i is a hyperparameter that reflects the weight of each corpus. It represents the contribution and importance of all participating corpora in the whole. When we can obtain that α is 1 through a large number of experiments, that is, when weights are not distinguished, the model reaches the highest performance, which is also consistent with the conclusion of Wang et al. [26].

This paper adopts the fully-shared mode, which means that all parameters of the model are completely shared except that a corresponding output layer is set for each corpus. We provide an auxiliary corpus for the main corpus. The fully shared multichannel BiGRU can capture shared feature representations for multiple corpora, which are fed into their respective output layers to generate prediction sequences.

3. Experimental Settings

In this section, the experimental settings are reported clearly, including optimizer and regularization, hyperparameters, corpora, and evaluation measures.

3.1. Optimizer and Regularization. Adam [42] (Adaptive Moment Estimation) is adopted as the optimizer of our model during training. It is an adaptive optimization method that dynamically updates the learning rate by computing the gradient's 1st moment estimate and 2nd moment estimate. Each adjusted learning rate is limited to a clear range, which ensures that the parameters are steadily updated.

We use dropout during model training to prevent overfitting. Dropout [43] is designed to randomly filter some hidden layer nodes according to the preset dropout rate so that they do not participate in the back propagation to update

parameters. The above operations can effectively prevent overfitting. They make the model more generalized.

3.2. Hyperparameters. Table 1 reports the experimental hyperparameter settings. The dimension based on the pre-trained word embeddings, character embeddings, and contextualized word embeddings is set to 200, 30, and 1024, respectively. We adopt the Adam to optimize our model during training. The dimension of GRU units is 100, and the dropout rate is 0.5. We set learning rate as 0.001, and the batch size is 32. In this paper, *BIO* labeling schema is employed to preprocess the original samples. *B* denotes the first token of entities in samples. *I* denotes the token located in entities. *O* denotes a token not belonging to entities.

3.3. Corpora. JNLPBA [44] and NCBI-disease [45] are our experimental main corpora. They are representative biomedical corpora of both multi and single classification. JNLPBA contains 5 types of entity: DNA, RNA, cell type, cell line, and protein. Training sets contain 2000 Medline abstracts, and test sets contain 404 Medline abstracts. The NCBI-disease corpus consists of 793 Medline abstracts, of which 593, 100, and 100, are used as training set, development set, and test set, respectively. It labels the disease name and the corresponding disease concept ID (the concept ID can be mapped to the ID in the MeSH or OMIM database). In addition, in the multitask learning framework, we use two other corpora as auxiliary data sets, namely BC2GM [46] and BC5CDR-disease [47]; the entity types contained in these two corpora are consistent with the main corpora. Table 2 provides the details of the above corpora.

3.4. Evaluation Measures. To evaluate the performance of our method, we adopt three conventional evaluation measures: precision (P), recall (R), and F1-score ($F1$). The calculation formulas are as follows:

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN} \tag{12}$$
$$F1 = \frac{2 * P * R}{P + R}$$

where TP denotes the number of true positive samples. TN denotes the number of true negative samples. FP denotes the number of false-positive samples. FN denotes the number of false-negative samples.

4. Results and Discussions

The described multitask learning multichannel BiGRU-CRF model is evaluated on NCBI-disease and JNLPBA. They are representative biomedical corpora of both single and multi-classification. We first compare the performance of each merge method and feature representations, as shown in Tables 3 and 4. Then, we evaluate the setting of hyperparameter values including the GRU dimension, optimizers, and dropout, as shown in Tables 5, 6, and 7. From Table 8, the effect of the CRF layer in our architecture is shown by an experiment. From Table 9, the effect of the multitask learning strategy is shown by an experiment. Lastly, the experiment compares the performance of multichannel BiGRU with some existing feature-based methods in BNER.

Table 1: Experimental parameter settings.

Hyperparameter	Value
Word dim	200
Char dim	30
ELMo dim	1024
GRU dim	100
Head	8
α_i	1
Dropout rate	0.5
Initial learning rate	0.001
Optimizer	Adam
Batch size	32
Labeling schema	BIO

4.1. Performance Comparison of Merge Methods. The merge methods affect the performance of capturing features. In the merge layer, inappropriate feature representations integration methods can result in information repetition and redundancy. It will have a negative impact on integrating information. Therefore, we evaluate the performance of different designing merge methods: addition, connection, unit-level attention, and channel-level attention. From Table 3, when the unit-level attention method is adopted, the model obtains the highest $F1$-Score. The probable reason is that the unit-level attention method can fully integrate the important features captured by each channel and do not interfere with each other; thus, we use the unit-level attention method in the merge layer.

4.2. Performance Comparison of each Representations. This paper proposes a multichannel BiGRU-CRF model to capture richer feature information by sending multiple representations individually or collectively into BiGRU. We evaluate the performance of each channel based on different representations while verifying the effectiveness of our multichannel method. The experimental results are shown in Table 4. It can be seen that the multichannel representations can provide richer potential information, and the concatenate representations are superior to the single representations. In summary, we compare the performance between each representation on the same corpus. Our merge-based multiple representations method achieves optimal performance, with the F1-scores of 76.0 and 88.7 on the JNLPBA and NCBI-disease corpora, respectively.

4.3. Performance Comparison of GRU Units Dimensions. GRU units' dimensions affect the ability of learning features and the performance of the classifier. Too few hidden units can result in insufficient capture features. Conversely, it may lead to information redundancy and increase the

Table 2: Introduction to experimental corpora.

Main	Entity types and counts	Size
NCBI-disease	Disease (6881)	793
JNLPBA	Gene/proteins (35336); cell line (4330); cell type (8649); DNA(10589); RNA(1069)	2404
Auxiliary	**Entity types and counts**	**Size**
BC5CDR-disease	Disease (12852)	1500
BC2GM	Gene/proteins (24583)	20000

Table 3: Performance comparison of the different merge methods.

Merge methods	JNLPBA			NCBI-disease		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Addition	72.9	78.8	75.7	87.4	88.6	88.0
Connection	71.1	78.3	74.5	85.3	89.2	87.2
Unit-level attention	**72.6**	**79.6**	**76.0**	**88.2**	**89.2**	**88.7**
Channel-level attention	72.1	78.6	75.2	86.6	88.7	87.6

Table 4: Performance comparison of each representations.

JNLPBA	Precision	Recall	F1-score	Δ	NCBI-disease	Precision	Recall	F1-score	Δ
Ours	**72.6**	**79.6**	**76.0**	—	**Ours**	**88.2**	**89.2**	**88.7**	—
ELMo	69.8	76.8	73.1	2.9	ELMo	83.8	85.4	84.6	4.1
Char	66.9	71.9	69.3	6.7	Char	83.7	80.6	82.1	6.6
Word	69.9	75.5	72.6	3.4	Word	84.2	80.9	82.5	6.2
ELMo+Char	71.5	76.4	73.8	2.2	ELMo+Char	86.0	85.5	85.7	3.0
Word+Char	68.9	77.6	73.0	3.0	Word+Char	84.2	85.1	84.7	4.0
Word+ELMo	70.1	77.5	73.6	2.4	Word+ELMo	84.5	86.0	85.3	3.4
Word+ELMo+Char	71.4	77.8	74.4	1.6	Word+ELMo+Char	87.2	85.9	86.6	2.1

Table 5: Performance comparison of GRU units' dimensions.

GRU	JNLPBA	Precision	Recall	F1-score	NCBI-disease	Precision	Recall	F1-score
Dimensions	50	70.7	77.5	73.9	50	87.0	87.5	87.2
	100	72.6	79.6	76.0	**100**	88.2	89.2	88.7
	150	70.3	77.3	73.6	150	88.1	85.7	86.9
	200	71.1	76.3	73.6	200	85.1	86.9	86.0

Table 6: Performance comparison of different optimization methods.

	Precision	Recall	F1-score
JNLPBA			
SGD	64.9	73.7	68.9
AdaGrad	72.0	75.0	73.4
Adam	72.6	79.6	76.0
NCBI-disease			
SGD	77.5	79.8	78.6
AdaGrad	85.4	86.5	85.9
Adam	88.2	89.2	88.7

computational burden. Both of them will have a negative impact on model performance. Therefore, we evaluate the performance of different neuron dimensions to obtain the best hyperparameters. We set the size of GRU units to be 50, 100, 150, 200 and evaluate them. As the results show in Table 5, when the dimensions are 100, it achieves the best performance. Therefore, the GRU units' dimensions are set to 100.

4.4. Performance Comparison of Combining CRF Layer. The CRF layer can capture the dependence between adjacent labels by transition probability. This paper evaluates the effectiveness of the CRF layer. The experimental results are shown in Table 8. After combining BiGRU with the CRF layer, the model performance has been significantly

Table 7: Performance comparison of using dropout.

	Precision	Recall	F1-score
JNLPBA			
No	73.5	72.9	73.2
Yes	72.6	79.6	76.0
Δ	**-0.9**	**+6.7**	**+2.8**
NCBI-disease			
No	86.9	84.8	85.9
Yes	88.2	89.2	88.7
Δ	**+1.3**	**+4.4**	**+2.8**

Table 8: Performance comparison of model with and without CRF layer.

	Precision	Recall	F1-score
JNLPBA			
BiGRU	70.8	73.7	72.3
BiGRU-CRF	72.6	79.6	76.0
Δ	**+1.8**	**+5.9**	**+3.7**
NCBI-disease			
BiGRU	82.4	86.7	84.5
BiGRU-CRF	88.2	89.2	88.7
Δ	**+5.8**	**+2.5**	**+4.2**

Table 9: Performance comparison of adopting multitask learning.

	Precision	Recall	F1-score
JNLPBA			
Single-task	71.8	79.7	75.6
Multi-task	72.6	79.6	76.0
Δ	**+0.8**	**-0.1**	**+0.4**
NCBI-disease			
Single-task	87.2	88.6	87.9
Multi-task	88.2	89.2	88.7
Δ	**+1.0**	**+0.6**	**+0.8**

improved on the JNLPBA and NCBI-disease corpora. It proves the validity of the CRF layer.

4.5. Performance Comparison of Adopting Multitask Learning. From the Table 9, the multitask learning strategy we adopted is effective. The auxiliary corpora provide more training samples and valuable information for the main corpora. According to the analysis of main corpora evaluation results, the multitask learning framework makes the performance improvement of JNLPBA less obvious than NCBI-disease. The possible reason is that the entity type of NCBI-disease is completely consistent with the auxiliary corpus BC5CDR-disease. The auxiliary corpus BC2GM contains only "protein", the training samples and relevant information of the other four entity types in the main corpus JNLPBA have not been supplemented.

4.6. Performance Comparison of Optimization Methods. The optimization method determines the convergence speed and performance of the model training process. This paper evaluates three different optimization methods: Adam, SGD, and AdaGrad. SGD is one of the commonly used optimizers during training. It randomly extracts fixed-size training samples to calculate gradients and update parameters. But it may lead to convergence to a local minimum. Compared to SGD, AdaGrad does not rely on a preset learning rate, but adaptively adjusts it during training. It is well suited to handle sparse data but may cause a vanishing gradient. The experimental results are shown in Table 6. Compared with the other two optimization methods, Adam achieves the fastest convergence speed and highest performance under the same conditions. Therefore, this paper uses Adam as the optimizer.

4.7. Performance Comparison of Using Dropout. This paper evaluates the effectiveness of dropout. The experimental results are shown in Table 7. After setting the dropout rate, the model performance has been significantly improved on the JNLPBA and NCBI-disease corpora. It demonstrates the validity of dropout.

4.8. Performance Comparison with Existing Feature-Based Methods. Lastly, we draw a comparison between our model and existing models. In order to ensure the fairness and rationality of the experiment, we have divided the existing models into two kinds according to the different training patterns. One kind is feature-based, which applies specific input representations to task-specific different architectures, such as the approaches listed in Table 10; while another kind is fine-tuning, which trains various downstream tasks with fine-tuning parameters in fixed model architectures, such as BERT [55]. This paper reports the performance comparison with existing models of feature-based representations.

The performance comparison results on the JNLPBA corpus are shown on the left side of Table 10. In these studies, the early methods (dictionary based and rule based) and the conventional machine learning models also obtained reasonable results in BNER, including Finkel et al. [1], Settles [3], Tsuruoka et al. [2], Tang et al. [6], Chang et al. [4], and Liao et al. [5]. NERBio [53] was the best rule-based system on a JNLPBA corpus, and the F1-score is 73.0. The Skip-Chain CRF adopted by Liao et al. [5] was the state-of-the-art conventional machine learning model. It obtained a reasonable F1-score of 73.2. Compared with the above best early method and conventional machine learning method, our model has increased F1-score values by 3.0 and 2.8, respectively. We can produce these results without any feature engineering but simple architecture. Compared with existing deep learning studies, the performance of our model is better than Li et al. [33]. They proposed a CNN-BLSTM-CRF model with word embeddings and character embeddings. Our model has increased the recall and F1-score by 9.7 and 1.6, respectively. Gridach et al. [54] proposed a BiLSTM-CRF model

Table 10: Performance comparison with existing feature-based methods.

Methods	JNLPBA				Methods	NCBI-disease			
	Type	*P*	*R*	*F*1		Type	*P*	*R*	*F*1
Finkel et al. [1]	S	71.6	68.6	70.1	Xu et al. [48]	S	84.8	76.1	80.2
Settles [3]	S	69.1	72.0	70.5	Leaman et al. [49]	S	82.8	81.9	80.9
Yao et al. [18]	S	64.9	76.1	71.0	Dogan et al. [45]	S	83.8	80.0	81.8
Tsuruoka et al. [2]	S	67.5	75.8	71.4	Leaman et al. [50]	S	85.1	80.8	82.9
Tang et al. [6]	S	70.8	72.0	71.4	Limsopatham et al. [22]	S	86.7	81.9	84.3
Chang et al. [4]	S	—	—	71.9	Wei et al. [51]	S	85.3	83.3	84.3
Zhu et al. [20]	S	—	—	72.6	Dang et al. [24]	S	85.0	83.8	84.4
Li et al. [21]	S	74.8	70.9	72.8	Habibi et al. [52]	S	86.4	82.9	84.6
Tsai et al. [53]	S	72.0	74.0	73.0	Zhao et al. [19]	S	85.1	85.3	85.2
Liao et al. [5]	S	72.8	73.6	73.2	Wang et al. [26]	M	85.9	86.4	86.1
Wang et al. [26]	M	70.9	76.3	73.5	Xu et al. [23]	S	86.6	85.8	86.2
Lyu et al. [25]	S	71.2	76.5	73.8	Yoon et al. [27]	M	85.5	87.3	86.4
Li et al. [33]	S	79.6	69.9	74.4	Zhu et al. [20]	S	86.5	88.1	87.3
Gridach et al. [54]	S	74.1	77.7	75.8	Sachan et al. [28]	T	86.4	88.3	87.3
Ours	**M**	**72.6**	**79.6**	**76.0**	**Ours**	**M**	**88.2**	**89.2**	**88.7**

∗"S" denotes the single-task model. "M" denotes the multitask model. "T" denotes the model based on transfer learning.

with pretrained word embeddings and character embeddings. They computed the character vectors by a bidirectional LSTM. This model significantly enhanced the best performance of single-task BNER models. The performance of our model is close to theirs. In summary, our method obtains promising results compared with existing feature-based models under the premise of using merge-based multiple features and simple architecture.

The performance comparison on the NCBI-disease corpus is shown in Table 10 (right side). In these studies, Leaman et al. [45, 49, 50] first adopted conventional machine learning methods to obtain competitive performance on the NCBI-disease dataset. They developed multiple BNER systems (e.g., DNorm and TaggerOne) in subsequent studies. The recent deep learning methods achieved satisfactory results in BNER. In addition to some of the related works described in the first section, including Limsopatham et al. [22], Dang et al. [24], Zhao et al. [19], Wang et al. [26], Xu et al. [23], Yoon et al. [27], Zhu et al. [20], and Sachan et al. [28], Xu et al. [48] proposed a three-layer neural network to identify disease entities. The BiLSTM with the same structure was used to generate character-level embeddings and capturing features. The entity labels were predicted through the CRF layer. Wei et al. [51] designed a hybrid model combining the conventional machine learning methods with neural networks, and bidirectional RNN and CRF were employed as submodels to extract features. Then, the output was merged and fed into SVM for classification. Habibi et al. [52] achieved reasonable performance on multiple biomedical datasets based on word embedding and a LSTM-CRF model. GRAM-CNN [20] was the best single-task system which was developed by CNN on the NCBI-disease corpus. It obtained an F1-score of 87.3. BiLM-NER [28] was the best feature-based model and was developed by the transfer learning method; the F1-score was 87.3. However, our model's performance is better than the above state-of-the-art work. Our model obtains the best performance among reported existing feature-based models.

4.9. Error Analysis. We analyze the error cases of the model on our corpora and summarized the main causes of these errors into the following two points.

The boundary is blurred. There are 3 main reasons for this error. First, biomedical entities are generally long and complex. For example, "Kappa B-specific DNA binding proteins" contains five words as the entity, and the length of entities in the general field is usually within three words. In addition, it contains the word "DNA", and the entity itself is "protein". Second, the virtual words and conjunctions within biomedical entities influence the judgment of the boundary. For example, there may be fixed-use conjunctions in biomedical entities, but they are often misjudged as "O". Finally, an entity in biomedical corpora is part of another entity, but they belong to two types. For example, "MZF-1" is part of "Recombinant MZF-1", but they belong to "DNA" and "protein". To a certain extent, these above issues are plaguing our model.

Corpora annotation inconsistency. For example, "wild-type" is labeled as "O" in "gave nearly wild-type levels of gene expression in phorbol ester-treated Jurkat cells but not in phorbol ester-treated HeLa or U937 cells.", but in "as a wild-type but not a mutant TSAP-binding site of the sea urchin functions only in transfected B cells as an upstream promoter element.", it is labeled as "DNA". In addition, there are abbreviations of entities in some biomedical sequences, and our model is difficult to identify. For example, "IL-2" in "Under the same conditions, Lck did not stimulate IL-2 promoter unless it was activated by mutation" and "Interleukin-2" in "The proteasome regulates receptor-mediated endocytosis of interleukin-2" refer to the same entity, but our model has difficulty to distinguish them.

These analyses demonstrate that the complexity and annotation inconsistency of biomedical corpora are major

factors that result in errors. To address these issues, we can disambiguate through entity linking during corpora preprocessing or adopt more external representations.

5. Conclusion

In this paper, we propose a multitask learning multichannel BiGRU-CRF model based on contextualized word representations. First, we obtain word, character, and contextualized word representations through a biomedical pretrained dictionary, convolutional neural networks, and ELMo pretrained language model, respectively. The character representations can describe the morphological features of words, and the contextualized word representations can better represent both polysemous and ambiguous words according to the context information. Then, we train multiple BiGRU submodels at the same time, each of which is viewed as a channel. The three representations are used as input for different channels, respectively, or in combination. Next, we design multiple methods to integrate the output of each channel in the merge layer. Finally, considering the dependence between labels, the CRF layer is adopted to parse sequence labels. It avoids outputting non-compliant label sequences. In addition, multitask learning strategy is adopted to solve the problem of insufficient training samples in specific fields. The auxiliary corpora with the same entity types are applied to supplement more training samples and relevant information for the main corpora to be evaluated. Our model has a simple architecture and avoids feature engineering. The multitask learning multichannel BiGRU-CRF achieves promising results on JNLPBA and NCBI-disease corpora, with F1-scores of 76.0 and 88.7, respectively. In the future, we plan to introduce more abundant additional features (e.g., domain knowledge base, structured ontology) to enhance the performance.

Data Availability

The data sets used in this paper are all publicly available. The related references of data sets adopted to support the findings of this study are included within this paper.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61976124).

References

[1] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting context for biomedical entity recognition: From syntax to the web," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 88–91, Geneva, Switzerland, 2004, Association for Computational Linguistics.

[2] Y. Tsuruoka, Y. Tateishi, J. D. Kim et al., "Developing a robust part-of-speech tagger for biomedical text," in *Panhellenic Conference on Informatics*, pp. 382–392, Volas, Greece, 2005, Springer.

[3] B. Settles, "Abner: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191-3192, 2005.

[4] F. Chang, J. Guo, W. Xu, and S. R. Chung, "Application of word embeddings in biomedical named entity recognition tasks," *Journal of Digital Information Management*, vol. 13, no. 5, 2015.

[5] Z. Liao and H. Wu, "Biomedical named entity recognition based on skip-chain crfs," in *2012 International Conference on Industrial Control and Electronics Engineering*, pp. 1495–1498, Xi'an, China, 2012.

[6] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu, "Evaluating word representation features in biomedical named entity recognition tasks," *BioMed Research International*, vol. 2014, Article ID 240403, 6 pages, 2014.

[7] K. Li, W. Ai, Z. Tang et al., "Hadoop recognition of biomedical named entity using conditional random fields," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 11, pp. 3040–3051, 2015.

[8] M. Chen and Y. Hao, "Label-less learning for emotion cognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 1–11, 2019.

[9] M. Chen, Y. Hao, H. Gharavi, and V. C. M. Leung, "Cognitive information measurements: a new perspective," *Information Sciences*, vol. 505, pp. 487–497, 2019.

[10] M. Chen, Y. Jiang, Y. Cao, and A. Y. Zomaya, "Creativebioman: Brain and body wearable computing based creative gaming system," 2019, https://arxiv.org/abs/1906.01801.

[11] M. Chen, Y. Jiang, N. Guizani et al., "Living with i-fabric: smart living powered by intelligent fabric and deep analytics," *IEEE Network*, pp. 1–8, 2020.

[12] Y. Chen, J. Tao, Q. Zhang et al., "Saliency detection via the improved hierarchical principal component analysis method," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8822777, 12 pages, 2020.

[13] Y. Chen, J. Wang, S. Liu et al., "Multiscale Fast Correlation Filtering Tracking Algorithm Based on a Feature Fusion Model," *Concurrency and Computation: Practice and Experience*, 2019.

[14] Y. Chen, W. Xu, J. Zuo, and K. Yang, "The fire recognition algorithm using dynamic feature fusion and iv-svm classifier," *Cluster Computing*, vol. 22, pp. 7665–7675, 2019.

[15] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive internet of things: improving sensitivity and interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58–64, 2019.

[16] Y. Zhang, Y. Qian, D. Wu, M. S. Hossain, A. Ghoneim, and M. Chen, "Emotion-aware multimedia systems security," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 617–624, 2019.

[17] Y. Zhang, R. Wang, M. S. Hossain, M. F. Alhamid, and M. Guizani, "Heterogeneous information network-based content caching in the internet of vehicles," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 10216–10226, 2019.

[18] L. Yao, H. Liu, Y. Liu, X. Li, and M. W. Anwar, "Biomedical named entity recognition based on deep neutral network," *International Journal of Hybrid Information Technology*, vol. 8, no. 8, pp. 279–288, 2015.

[19] Z. Zhao, Z. Yang, L. Luo et al., "Disease named entity recognition from biomedical literature using a novel convolutional neural network," *BMC Medical Genomics*, vol. 10, no. S5, 2017.

[20] Q. Zhu, X. Li, A. Conesa, and C. Pereira, "GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text," *Bioinformatics*, vol. 34, no. 9, pp. 1547–1554, 2018.

[21] L. Li, L. Jin, Y. Jiang, and D. Huang, "Recognizing biomedical named entities based on the sentence vector/twin word embeddings conditioned bidirectional lstm," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pp. 165–176, Springer, 2016.

[22] N. Limsopatham and N. Collier, "Learning orthographic features in bi-directional lstm for biomedical named entity recognition," in *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016)*, pp. 10–19, Osaka, Japan, 2016.

[23] K. Xu, Z. Zhou, T. Gong, T. Hao, and W. Liu, "SBLC: a hybrid model for disease named entity recognition based on semantic bidirectional LSTMs and conditional random fields," *BMC Medical Informatics and Decision Making*, vol. 18, no. S5, 2018.

[24] T. H. Dang, H. Q. Le, T. M. Nguyen, and S. T. Vu, "D_3ner: biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information," *Bioinformatics*, vol. 34, no. 20, pp. 3539–3546, 2018.

[25] C. Lyu, B. Chen, Y. Ren, and D. Ji, "Long short-term memory RNN for biomedical named entity recognition," *BMC Bioinformatics*, vol. 18, no. 1, 2017.

[26] X. Wang, Y. Zhang, X. Ren et al., "Cross-type biomedical named entity recognition with deep multi-task learning," 2018, https://arxiv.org/abs/1801.09851.

[27] W. Yoon, C. H. So, J. Lee, and J. Kang, "Collabonet: collaboration of deep neural networks for biomedical named entity recognition," 2018, https://arxiv.org/abs/1809.07950.

[28] D. S. Sachan, P. Xie, M. Sachan, and E. P. Xing, "Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition," 2017, https://arxiv.org/abs/1711.07908.

[29] M. E. Peters, M. Neumann, M. Iyyer et al., "Deep contextualized word representations," 2018, https://arxiv.org/abs/1802.05365.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, https://arxiv.org/abs/1301.3781.

[31] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*, pp. 1532–1543, Doha, Qatar, 2014.

[32] C. Quan, L. Hua, X. Sun, and W. Bai, "Multichannel convolutional neural network for biological relation extraction," *BioMed Research International*, vol. 2016, Article ID 1850404, 10 pages, 2016.

[33] L. Li and Y. Guo, "Biomedical named entity recognition with cnn-blstm-crf," *Journal of Chinese Information Processing*, vol. 32, no. 1, pp. 116–122, 2018.

[34] D. Zeng, C. Sun, L. Lin, and B. Liu, "Lstm-crf for drug-named entity recognition," *Entropy*, vol. 19, no. 6, 2017.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, https://arxiv.org/abs/1409.1259.

[37] K. Cho, B. Van Merriënboer, C. Gulcehre et al., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014, https://arxiv.org/abs/1406.1078.

[38] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, https://arxiv.org/abs/1412.3555.

[39] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[40] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *International conference on machine learning*, pp. 2342–2350, Lille, France, 2015.

[41] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, https://arxiv.org/abs/1603.01360.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[44] J. D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04*, pp. 70–75, Geneva, Switzerland, 2004.

[45] R. I. Doğan, R. Leaman, and Z. Lu, "Ncbi disease corpus: a resource for disease name recognition and concept normalization," *Journal of Biomedical Informatics*, vol. 47, pp. 1–10, 2014.

[46] L. Smith, L. K. Tanabe, R. J. n. Ando et al., "Overview of biocreative ii gene mention recognition," *Genome Biology*, vol. 9, no. S2, 2008.

[47] C. H. Wei, Y. Peng, R. Leaman et al., "Overview of the biocreative v chemical disease relation (cdr) task," in *Proceedings of the fifth BioCreative challenge evaluation workshop*, vol. 14, Seville, Spain, 2015.

[48] K. Xu, Z. Zhou, T. Hao, and W. Liu, "A bidirectional lstm and conditional random fields approach to medical named entity recognition," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017*, pp. 355–365, Cairo, Egypt, 2018, Springer.

[49] R. Leaman, R. I. Dogan, and Z. Lu, "Dnorm: disease name normalization with pairwise learning to rank," *Bioinformatics*, vol. 29, no. 22, pp. 2909–2917, 2013.

[50] R. Leaman and Z. Lu, "Taggerone: joint named entity recognition and normalization with semi-markov models," *Bioinformatics*, vol. 32, no. 18, pp. 2839–2846, 2016.

[51] Q. Wei, T. Chen, R. Xu, Y. He, and L. Gui, "Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks," *Database*, vol. 2016, article baw140, 2016.

[52] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, 2017.

[53] R. Tsai, C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung, and W.-L. Hsu, "NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition," *BMC Bioinformatics*, vol. 7, Supplementary 5, 2006.

[54] M. Gridach, "Character-level neural network for biomedical named entity recognition," *Journal of Biomedical Informatics*, vol. 70, pp. 85–91, 2017.

[55] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, https://arxiv.org/abs/1810.04805.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 9595687, 13 pages
https://doi.org/10.1155/2020/9595687

Research Article

A Multiscale-Based Adjustable Convolutional Neural Network for Multiple Organ Segmentation

Zhiqiang Tian,[1] Jingyi Song,[1] Chenyang Zhang,[1] Xiaohui Tian,[1] Zhong Shi,[2,3] and Xiaofu Yu[2,3]

[1]*School of Software Engineering, Xi'an Jiaotong University, 710049, China*
[2]*Institute of Cancer and Basic Medicine (ICBM), Chinese Academy of Sciences, 310022, China*
[3]*Cancer Hospital of the University of Chinese Academy of Sciences, 310022, China*

Correspondence should be addressed to Zhiqiang Tian; zhiqiangtian@xjtu.edu.cn

Received 7 February 2020; Revised 16 June 2020; Accepted 21 July 2020; Published 3 August 2020

Academic Editor: Yin Zhang

Accurate segmentation ofs organs-at-risk (OARs) in computed tomography (CT) is the key to planning treatment in radiation therapy (RT). Manually delineating OARs over hundreds of images of a typical CT scan can be time-consuming and error-prone. Deep convolutional neural networks with specific structures like U-Net have been proven effective for medical image segmentation. In this work, we propose an end-to-end deep neural network for multiorgan segmentation with higher accuracy and lower complexity. Compared with several state-of-the-art methods, the proposed accuracy-complexity adjustment module (ACAM) can increase segmentation accuracy and reduce the model complexity and memory usage simultaneously. An attention-based multiscale aggregation module (MAM) is also proposed for further improvement. Experiment results on chest CT datasets show that the proposed network achieves competitive Dice similarity coefficient results with fewer float-point operations (FLOPs) for multiple organs, which outperforms several state-of-the-art methods.

1. Introduction

Radiation therapy is the main clinical method of treating various cancers [30]. It can be seen as a trade-off between sending maximum dose to the target-volume (TV) and minimum dose to the OARs [1, 31]. Accurate segmentation of OARs contributes to the protection of normal organs during treatment [32]. Manually delineating OARs slice-by-slice in CT scans requires expertise and lots of time. Adjacent soft tissue with low contrast and noise blurs the boundaries of organs, which may also lead to delineation errors. Therefore, automatic segmentation has become a research hotspot [33]. Such a framework could help radiologists segment OARs more accurately with much less time.

Deep convolutional models have shown state-of-the-art performance in image segmentation especially after fully convolutional networks (FCN) [2] were proposed. In FCN, fully convolutional layers are substituted for fully connected layers in a classic classification network to reserve spatial information, which makes FCN a reasonable choice for dense prediction. Based on FCN, Badrinarayanan et al. [3] introduced a concept of encoders and decoders and upsampled the low-level inputs in turn to the original resolution, which achieves impressive performance in scene understanding applications. U-Net [4] adopted the similar encode and decode paths and used several skip connections between them to combine the low-level and high-level features. Thanks to this carefully designed structure, it achieves considerable success in medical image segmentation tasks. In addition, great efforts have been subsequently made to improve the performance of U-Net [5, 6].

Targets in images usually have various sizes. It is crucial to let a network "see" objects of different scales. What kind of multiscale information to use and how to integrate information are the purpose of this work. PSPNet [7] merged convolutional features from different region-based context

aggregations. DeepLabs [8–11] similarly introduced a spatial pyramid pooling with different dilated kernels. An attention mechanism was first proposed to model long-range dependencies in machine translation and has been successfully transferred to computer vision tasks. It allocates more weights to the potentially interesting regions and suppresses the less informative ones. DANet [12] built a dual attention module to simultaneously model channel-wise and spatial-wise semantics. Oktay et al. [5] proposed an attention gate in the skip connection layer of the U-Net to teach the decoder where to "look." Selective Kernel [13] (SK) took advantage of the attention to guide the fusion of multiscale information. It embeds multiple lightweight SE [14] attentions to dynamically select the receptive field size of each neuron in convolutional layers, which is also consistent with the neuroscience cognition that the visual cortical neuron adaptively adjusts the size of its receptive field according to stimulus. In such case, the learned attention maps play the role of a stimulus.

Aside from accuracy, computation complexity is another factor to measure model performance. A widely used evaluation metric of computation complexity is the number of float-point operations, also known as FLOPs [15]. Substantial efforts have been made to reduce FLOPs of convolutional networks. MobileNets [16, 17] adopted depth-wise separable convolution, which splits a standard convolution into a depth-wise and a point-wise convolution. It can greatly reduce the computation cost. ShuffleNets [18, 19] further integrated group convolution and reduced channel redundancy in point-wise connectivity. These methods focus on sparse connection between channels. In contrast, Octave Convolution [20] (OctConv) focused on the spatial dimension and argued that potential redundancy may exist. Each location independently stores its own features, ignoring the possibility that there could be similar information between adjacent locations. OctConv presented a multifrequency processing algorithm, which splits feature maps into high and low frequency to reduce spatial redundancy. Through sharing information with adjacent neurons, the spatial resolution of the low-frequency parts can be reduced, which helps to save computation and memory cost.

A highly generalized CNN should take the computing power and memory consumption into consideration or otherwise may fail to run on the limited computation conditions, especially the training stage [21]. Several strategies have been proposed to alleviate this issue. A most straightforward method is to reduce batch size during training. Heavy models like V-Net [22] and No-New-Net [23] adopted a small batch size of two. However, reducing batch size would make the network hard to converge. In particular, when the batch size is reduced to one, the Batch Normalization [24] that is important for stable gradient propagation becomes invalid. Another possible way to reduce computation is to compress the width and depth of the network and image resolution, which is under the risk of accuracy decrease. Researches have shown that the increase of any of these three elements can improve model performance (the width [25], the depth [26], the image size [27], and all three aspects [28]). In this paper, we propose an end-to-end network with symmetrical encode and decode paths to automatically segment multiple organs. The model complexity of the proposed network can be flexibly adjusted to meet the computation requirements without accuracy sacrifice. Neither the batch size nor the model capacity needs to be changed. Concretely, we introduce an accuracy-complexity adjustment module (ACAM) inspired by Chen et al. [20] throughout the encoder and decoder. A hyperparameter is adopted to control the compression of spatial redundancy of feature maps. To further enrich feature representations and capture information of organs in different sizes, we add a nonlinear multiscale aggregation module (MAM) after the encoder. The branches of the module have different rates of dilations, which is inspired by Li et al. [13].

Our contributions are summarized as follows:

(i) We introduce an accuracy-complexity adjustment module throughout the encoder and decoder to increase the segmentation accuracy and reduce the model complexity and memory usage simultaneously

(ii) We present an attention-based multiscale aggregation module after the encoder to enrich feature representation for further boosting the segmentation accuracy

(iii) The proposed network achieves competitive results, which outperform several state-of-the-art methods on chest CT segmentation datasets

The rest of this paper is organized as follows. In Section 2, the proposed method is described in detail. In Section 3, experimental results of our method and the state-of-the-art methods are presented. Finally, we state the conclusion in Section 4.

2. Materials and Methods

Our network is a U-shaped architecture, which contains two symmetrical encode and decode paths [4]. Skip connections are used to transfer the features from the encoder directly to the decoder. ACAMs are applied all-through the encoder and decoder. MAM is added after the encoder to further enrich feature representation and boost segmentation performance. Details of the network architecture can be seen in Figure 1. The numbers of channels of all layers can be seen in Table 1.

2.1. Accuracy-Complexity Adjustment Module. Since images can be divided into high frequency with details and low frequency with global layout, Chen et al. [20] infer that feature maps can also be divided into high and low frequency with different spatial contexts. Following this idea, we present an ACAM and apply it in every convolutional layer to get high- and low-frequency feature maps. The width and the height of the low frequency are reduced to half size, and the number of channels is controlled by a parameter α to meet the computation resource.

To be specific, an input feature map $X \in \mathscr{R}^{c\times h\times w}$, where $h \times w$ denotes the spatial resolution and c the number of

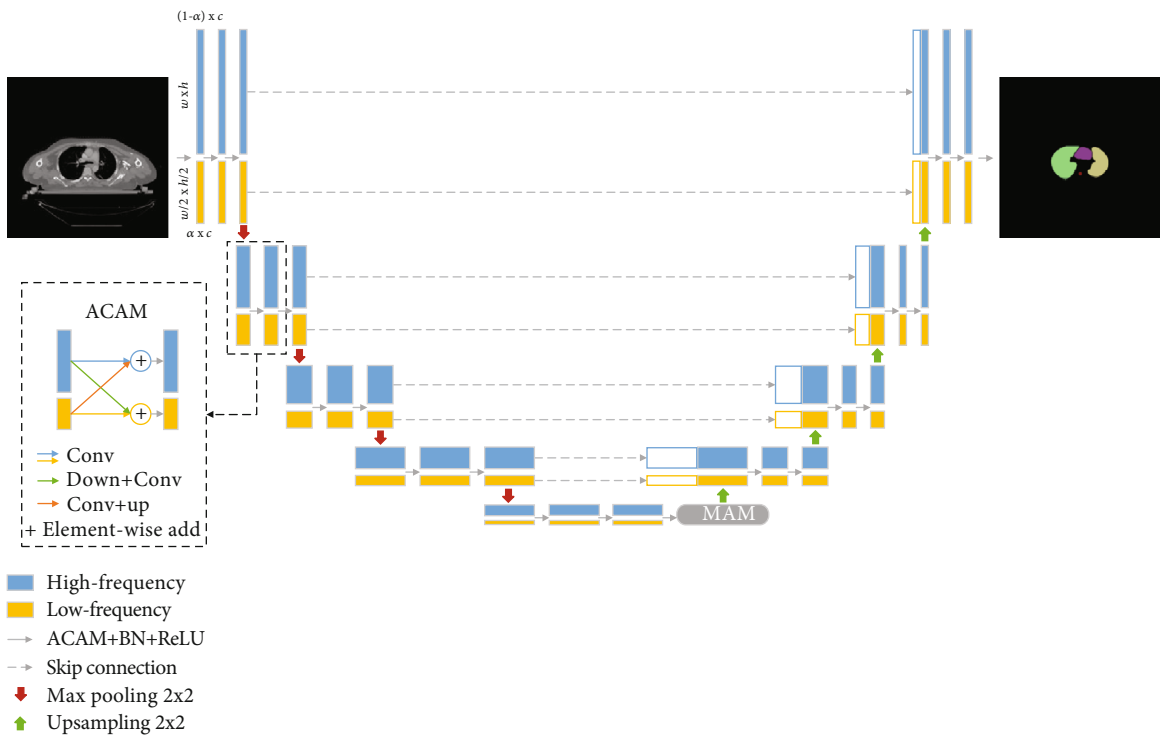


Figure 1: The architecture of the proposed network. The accuracy-complexity adjustment module (ACAM) is applied all-through the encoder and decoder. $w \times h$ and c represent the resolution and channels of the feature map, respectively. α is a hyperparameter controlling the ratio of the low frequency along the channel dimension. The multiscale aggregation module (MAM) is added after the encoder to enrich feature representation.

Table 1: The numbers of channels of all layers.

Layer name	Input channels		Output channels	
	High frequency	Low frequency	High frequency	Low frequency
InConv	1	—	$64(1-\alpha)$	64α
DownConv1	$64(1-\alpha)$	64α	$128(1-\alpha)$	128α
DownConv2	$128(1-\alpha)$	128α	$256(1-\alpha)$	256α
DownConv3	$256(1-\alpha)$	256α	$512(1-\alpha)$	512α
DownConv4	$512(1-\alpha)$	512α	$512(1-\alpha)$	512α
MAM	$512(1-\alpha)$	512α	$512(1-\alpha)$	512α
UpConv1	$1024(1-\alpha)$	1024α	$256(1-\alpha)$	256α
UpConv2	$512(1-\alpha)$	512α	$128(1-\alpha)$	128α
UpConv3	$256(1-\alpha)$	256α	$64(1-\alpha)$	64α
UpConv4	$128(1-\alpha)$	128α	$64(1-\alpha)$	64α
OutConv	$64(1-\alpha)$	64α	num of classes	—

channels, can be factorized into $X = \{X^H, X^L\}$ along the channel dimension via α. $X^H \in \mathscr{R}^{(1-\alpha)c\times h\times w}$ and $X^L \in \mathscr{R}^{\alpha c\times h/2\times w/2}$ represent high and low frequency, respectively, which is shown in Figure 2. These two frequency feature maps extract information independently through intrafrequency update and communicate with each other through interfrequency exchange. An output feature map is represented as $Y = \{Y^H, Y^L\}$, which can be defined as follows:

$$Y^H = f^{H\to H}(X^H) + f^{L\to H}(X^L), \tag{1}$$

$$Y^L = f^{L\to L}(X^L) + f^{H\to L}(X^H), \tag{2}$$

where $f^{H\to H}$ and $f^{L\to L}$ denote intraupdates, while $f^{L\to H}$ and $f^{H\to L}$ denote interexchanges. The intra- and interfrequency communications help to strengthen information propagation between channels, which brings potential improvements.

To compute the output feature map Y, a standard convolution kernel with weight W of $k \times k$ size is split into four parts $W = \{W^{H\to H}, W^{L\to H}, W^{L\to L}, W^{H\to L}\}$ via α shown in Figure 3. For the intrapart, a vanilla convolution is the only requirement, while for the interpart, resolution matching (i.e., downsampling or upsampling) should be performed before or after convolution. Here, we use average pooling for downsampling and nearest interpolation for upsampling, which is shown in the bottom left corner of Figure 1. Then, Equation (1) can be rewritten as

$$Y^H = f(X^H; W^{H\to H}) + \text{interpolate}(f(X^L; W^{L\to H}), 2), \tag{3}$$

$$Y^L = f(X^L; W^{L\to L}) + f(\text{pool}(X^H, 2); W^{H\to L}), \tag{4}$$

where $f(X; W)$ denotes a convolution function with weight parameter W, interpolate$(*, m)$ denotes upsampling by a

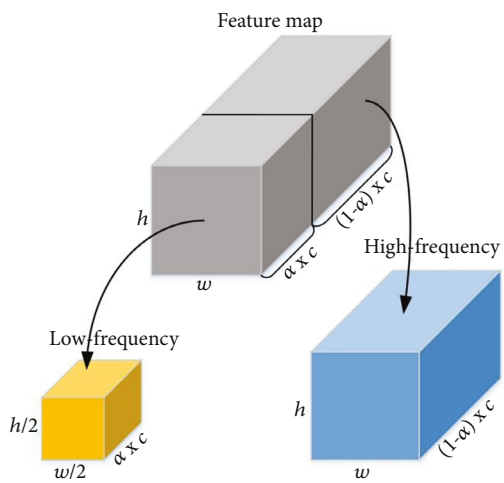


Figure 2: Representation of high- and low-frequency feature maps. The ratio α is a hyperparameter of the ACAM to control the number of channels of high- and low-frequency feature maps. Concretely, $c_{\text{low}} = \alpha \times c$ and $c_{\text{high}} = (1-\alpha) \times c$, where c is the total number of channels.

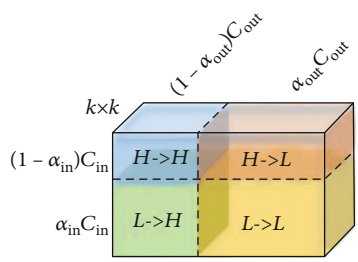


Figure 3: The $k \times k$ sized convolution kernel of accuracy-complexity adjustment module (ACAM).

factor of m, and pool$(*,n)$ denotes downsampling with kernel size $n \times n$ and strike n.

In ACAM, two operation strategies are used to improve the performance. First, the upsampling is performed after convolution in Y^H, and downsampling is performed before convolution in Y^L. Such an order makes convolutions operate on smaller feature maps, which can save computations. Second, using the same size of $k \times k$ kernels on the low frequency is equivalent to enlarging the receptive field size in the original pixel space, which introduces additional multiscale information.

The setting of α can be divided into three stages. In the first convolutional layer, we set $\alpha_{\text{in}} = 0$ and $\alpha_{\text{out}} = \alpha$, in which case X^L is disabled. This stage is usually used at the beginning of the network, where an original CT image is split into high and low frequency. The high frequency is obtained by a convolution function, and the low frequency is obtained through a sequential downsampling and convolution operation. In the last convolutional layer, we set $\alpha_{\text{in}} = \alpha$ and $\alpha_{\text{out}} = 0$, in which case Y^L is disabled. This stage is usually used at the end of the network, mapping the feature map to the prediction mask. The low frequency is restored to high frequency through a sequential convolution and an upsampling operation, which is added to the high-frequency part to get the final result. In all the middle convolutional layers, we keep $\alpha_{\text{in}} = \alpha_{\text{out}} = \alpha$ in ACAMs. The intermediate feature maps are obtained by Equation (3).

There are three kinds of operations in the ACAM, which are convolution, pooling, and interpolation. Compared with the convolution, the FLOPs of the pooling and interpolation are negligible. Therefore, we only compute the FLOPs of the convolution. For each convolution operation in ACAM, namely, $f^{H\to H}$, $f^{L\to H}$, $f^{L\to L}$, and $f^{H\to L}$, the theoretical FLOPs can be obtained by

$$\begin{aligned}
\text{FLOPs}\left(f^{H\to H}\right) &= h \times w \times k^2 \times (1-\alpha)^2 \times c^2,\\
\text{FLOPs}\left(f^{L\to H}\right) &= \frac{h}{2} \times \frac{w}{2} \times k^2 \times (1-\alpha) \times \alpha \times c^2,\\
\text{FLOPs}\left(f^{L\to L}\right) &= \frac{h}{2} \times \frac{w}{2} \times k^2 \times \alpha^2 \times c^2,\\
\text{FLOPs}\left(f^{H\to L}\right) &= \frac{h}{2} \times \frac{w}{2} \times k^2 \times \alpha \times (1-\alpha) \times c^2.
\end{aligned} \tag{5}$$

The final FLOPs include four subconvolutions, which are merged as follows:

$$\text{FLOPs} = \left(1 - \frac{3}{4}\alpha(2-\alpha)\right) \times h \times w \times k^2 \times c^2. \tag{6}$$

Therefore, by controlling the rate of low-frequency α, ACAM can save $(3/4)\alpha(2-\alpha)$ computations compared with a vanilla convolution.

2.2. Multiscale Aggregation Module. Targets in medical images usually have different sizes (e.g., spinal cord and lungs in this segmentation task). The size of a specific organ also varies in different slices. Multiscale information should be taken into consideration during training. Therefore, the MAM is added after the encoder to enrich the feature representation.

Formally, let $X \in \mathscr{R}^{c\times h\times w}$ be an input feature map, where $h \times w$ denotes the spatial resolution and c the number of channels. The MAM with M branches is proposed for X to integrate multiscale information. Taking $M = 4$ as an example, the module structure is shown in Figure 4. The basic 3×3 sized convolution kernels corresponding to different branches have different dilated rates. The MAM generates M feature maps in parallel, namely, $U = [U_1, U_2, \cdots, U_M]$. Remember that the purpose of MAM is to adaptively adjust the receptive field sizes of neurons in the next layer with multiscale information from these branches. Therefore, an element-wise fusion operation needs to be done first to integrate all the learned feature maps:

$$U = \sum_{b=1}^{M} U_b. \tag{7}$$

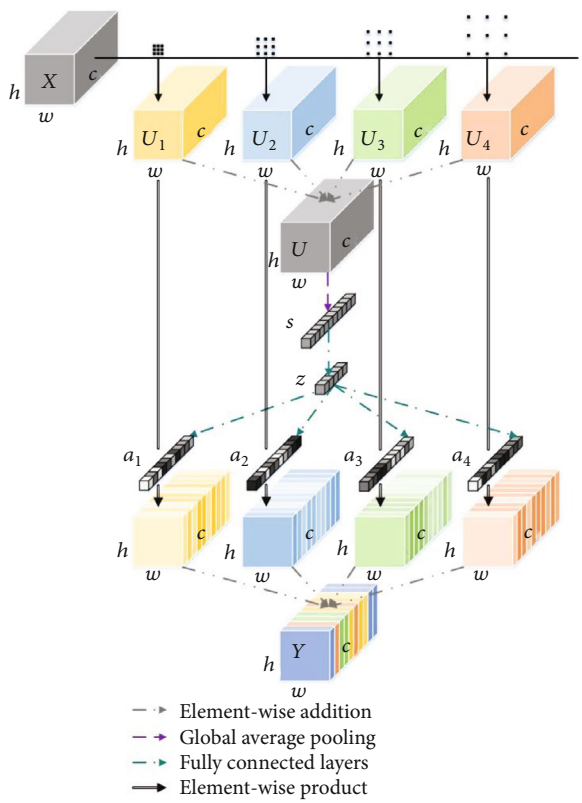


Figure 4: Multiscale aggregation module. For simplicity, we keep the number of channels c of $[U_1, U_2, \cdots, U_M]$ the same as X.

Then, the channel-wise statistics of U, expressed as a tensor $s \in \mathscr{R}^c$, can be built through a global average pooling to embed spatial context of each channel:

$$s = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} U(i, j). \tag{8}$$

Then, s is passed through a bottleneck made by a fully connected layer with weight $W_1 \in \mathscr{R}^{c \times d}$ and transformed into a compressed tensor $z \in \mathscr{R}^d$, where d is

$$d = \max\left(\frac{c}{r}, L\right). \tag{9}$$

The ratio r relates to the degree of compression, which helps to limit model complexity and assist feature generalization. L denotes a safe margin, which is used to prevent too much loss of channel-wise information caused by overcompression.

In order to separately calculate the attentions of M branches, z is split into M independent tensors $[a_1, a_2, \cdots, a_M]$ through fully connected layers with weight $W_2 \in \mathscr{R}^{M \times d \times c}$, decoding back to the original size. Then, an activation function $\sigma(\cdot)$ across c channels is used to build the corresponding attention maps $[\sigma(a_1), \sigma(a_2), \cdots, \sigma(a_M)]$, which is a softmax function in this case. These attention maps emphasize the important channels and ignore the less important ones in branches, acting as gates to constraint the information propagation. The final output Y is obtained through a summation over channel-wise products of these attention weights and feature maps:

$$Y = \sum_{i=1}^{M} a_i \times U_i. \tag{10}$$

Table 2: The effect of α on the performance. We calculate the percentage of FLOPs and memory relative to the baseline U-Net (the first row).

α	DSC (%)	FLOPs (G)	Memory (10^6)
Baseline ($\alpha = 0$)	94.34	123.93	1439
0.125	95.43	102.09 (82%)	1313 (91%)
0.25	95.41	83.26 (67%)	1188 (83%)
0.5	95.32	54.28 (44%)	938 (65%)
0.75	95.09	36.84 (30%)	687 (47%)
0.875	94.88	32.45 (26%)	562 (39%)
1.0	93.52	30.98 (25%)	365 (25%)

There are four operations in one MAM, which are convolutions, global average poolings, fully connected layers, and softmax. To analyze the model complexity, we compute FLOPs of each operation individually. There are $M + 1$ fully connected layers, which are all performed between one-dimensional tensors. Their total FLOPs are at most $(M + 1) \times c^2$, depending on the ratio parameter r, which is negligible compared with convolutions. We also ignore the FLOPs of softmax operation, which is performed M times on one-dimensional tensors. Element-wise additions and multiplies are conducted $2(M - 1)$ and M times, respectively. The sum of their FLOPs is $(3M - 2) \times h \times w \times c$, also much smaller than that of convolutions. Therefore, the main additional complexity is brought by the convolution performed early in each branch. To reduce the computation cost of the MAM, we put it in a bottleneck layer and compress the number of channels through point-wise convolutions.

3. Results and Discussion

3.1. Dataset and Metrics. An in-house dataset was used to evaluate the proposed method, which is provided by the Institute of Cancer and Basic Medicine (ICBM), Chinese Academy of Sciences, and Cancer Hospital of the University of Chinese Academy of Sciences. Manually labelled masks of the left lung, right lung, heart, and spinal cord of 36 patients were defined by two experienced radiation oncologists. The CT scans have 512×512 pixels in resolution. The pixel spacing varies from 0.95 mm to 1.37 mm, and all thicknesses are 5 mm. We randomly shuffled the dataset and split it into three subsets: training set, validation set, and test set. The training set accounts for 70% of the entire data set. The validation set accounts for 10%, and the test set accounts for 20%. No preprocessing was performed before training. A public dataset called SegTHOR is also used to test our method, which can be found in CodaLab. There are four organs in this dataset, esophagus, heart, trachea, and aorta. The CT scans have 512×512 pixels in resolution. The pixel

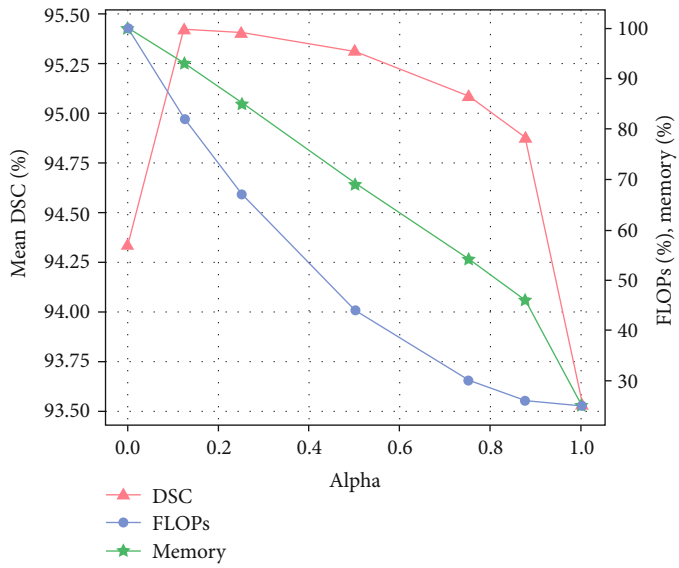


Figure 5: The relationships between DSC, FLOPs, memory, and α. Since the actual values of FLOPS and memory correspond to different scales, we only plot their percentages relative to the baseline.

spacing varies from 0.90 mm to 1.37 mm. The thicknesses vary from 2 mm to 3.7 mm. Image data and labels of 40 patients can be downloaded to train the network. Image data of 20 patients can be tested offline and uploaded onto the website to obtain the test results (https://competitions.codalab.org/competitions/21145). In the following sections, except for special declarations, we default to using the in-house dataset for experiments.

The proposed method was implemented by using a deep learning framework called Pytorch and trained on a single NVIDIA GTX 1080 Ti GPU with 11 GB of memory. The code of our complete method can be found online https://github.com/Jennsoo/UNet-Accuracy-Complexity.git. The network was trained with a stochastic gradient descend optimizer with an initial learning rate 0.03, momentum 0.9, and weight decay 0.0005. Poly learning rate policy was adopted to further decrease the learning rate by a factor of $(1-(\text{iter}/\text{total}_{\text{iters}}))^{0.9}$. We trained our model using crossentropy loss with 100 epochs in total. Weights were initialized according to MSRA [29], a zero-mean Gaussian distribution with variance $2/n$, in which n is the number of input elements. The Dice similarity coefficient (DSC) is our evaluation metric, which is defined as follows:where X is the ground truth and Y is the prediction.

3.2. Guidance for Adjusting Accuracy and Complexity. The hyperparameter α in ACAM is the main factor to balance the accuracy and model complexity. To evaluate how α affects the DSC-FLOP trade-off in the segmentation task, we conduct an experiment with $\alpha \in \{0.0, 0.125, 0.25, 0.5, 0.75, 0.875, 1.0\}$ on our in-house dataset. The corresponding mean DSC, FLOPs, and memory are listed in Table 2. In particular, $\alpha = 0$ relates to the baseline, which means no low frequency is separated out. We observe that ACAMs notably improve the segmentation DSC, while having less computation complexity and memory cost. Such boost can be attributed to the multifrequency processing in ACAMs, which brings an extra multiscale complement. We plot the mean DSC, FLOPs, and memory in Figure 5. From the figure, we can see that the DSC first significantly increases and then slowly declines, while FLOPs and memory decrease monotonically. The network reaches its best DSC at $\alpha = 0.125$, about 1% DSC improvement, while decreasing 18% complexity and 9% memory cost. For higher α, the DSC slowly goes down from the peak but is still better than the baseline. An intuitive explanation is that the spatial compression of low-frequency feature maps does not lead to serious information loss. This confirms the argumentation that feature maps in convolutional layers do have spatial redundancy. Through compressing the feature maps, computation and memory cost can be effectively saved without sacrificing accuracy. This experiment provides guidance for α selection. The DSC of $\alpha = 0.25$ is almost the same with the DSC of $\alpha = 0.125$, but the FLOPs and memory of $\alpha = 0.25$ take more advantages. Therefore, we set $\alpha = 0.25$ for all the ACAMs.

Table 3: The effect of the number of branches. We calculate the increased percentage of FLOPs and memory relative to the baseline U-Net (the first row).

#Branches (dilation)	DSC (%)	FLOPs (G)	Memory (10^6)
Baseline	94.34	123.93	1439
2 (1, 2)	94.54	125.41 (+1.19%)	1454 (+1.04%)
3 (1, 2, 4)	94.71	126.02 (+1.69%)	1457 (+1.25%)
4 (1, 2, 4, 8)	95.02	126.62 (+2.17%)	1460 (+1.46%)

3.3. The Influence of the Number of Branches in Multiscale Aggregation Module. Here, we conduct an experiment on the in-house dataset to explore the best choice of branches of the MAM. Concretely, we use kernels with different dilated rates to generate feature maps with different receptive field sizes. The basic kernel size is 3×3 with dilated rate 1. For each new branch, the dilated rate of the kernel is multiplied

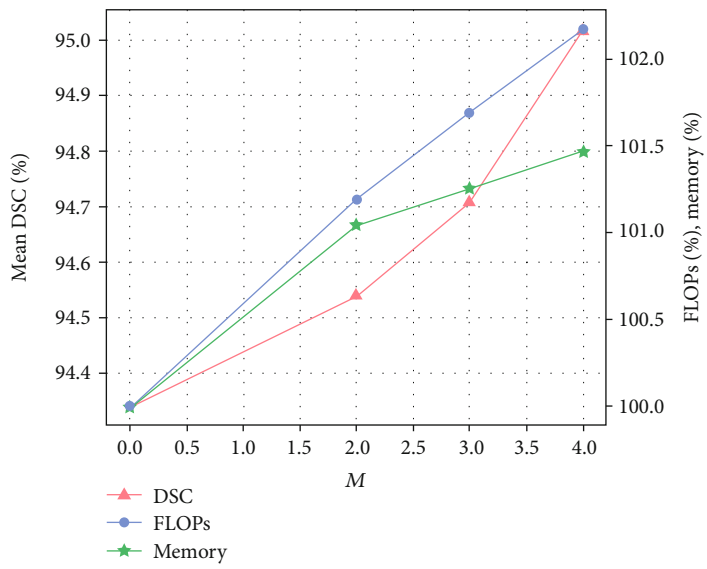


Figure 6: The relationships between DSC, FLOPs, memory, and the number of branches M.

Table 4: The effect of the number of branches. We calculate the increased percentage of FLOPs and memory relative to the baseline U-Net (the first row).

	L-lung	R-lung	Heart	Sp-Co	Mean
Baseline	96.86	96.88	94.40	89.20	94.34
Baseline+MAM	97.53	97.27	94.97	90.31	95.02
Baseline+ACAMs	97.63	97.43	95.20	91.36	95.41
Baseline+ACAMs+MAM	97.76	97.52	95.27	92.24	95.70

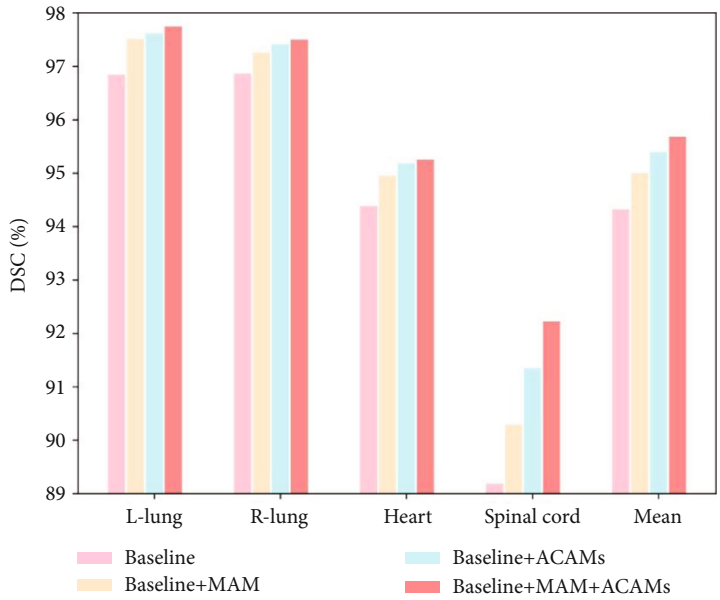


Figure 7: The DSC of each single organ of the ablation experiment.

by a power of two. Note that a standard MAM requires at least two branches. In this paper, three designs of branches are presented, and its corresponding results are recorded in Table 3. It is worth mentioning that we have tried to add the MAM in every convolutional layer as a parallel module, but the results were not satisfactory. We plot the mean DSC, FLOPs, and memory in Figure 6. No preprocessing was performed before training.

We observe that DSC is positively correlated with FLOPs and memory. The MAM with more branches performs

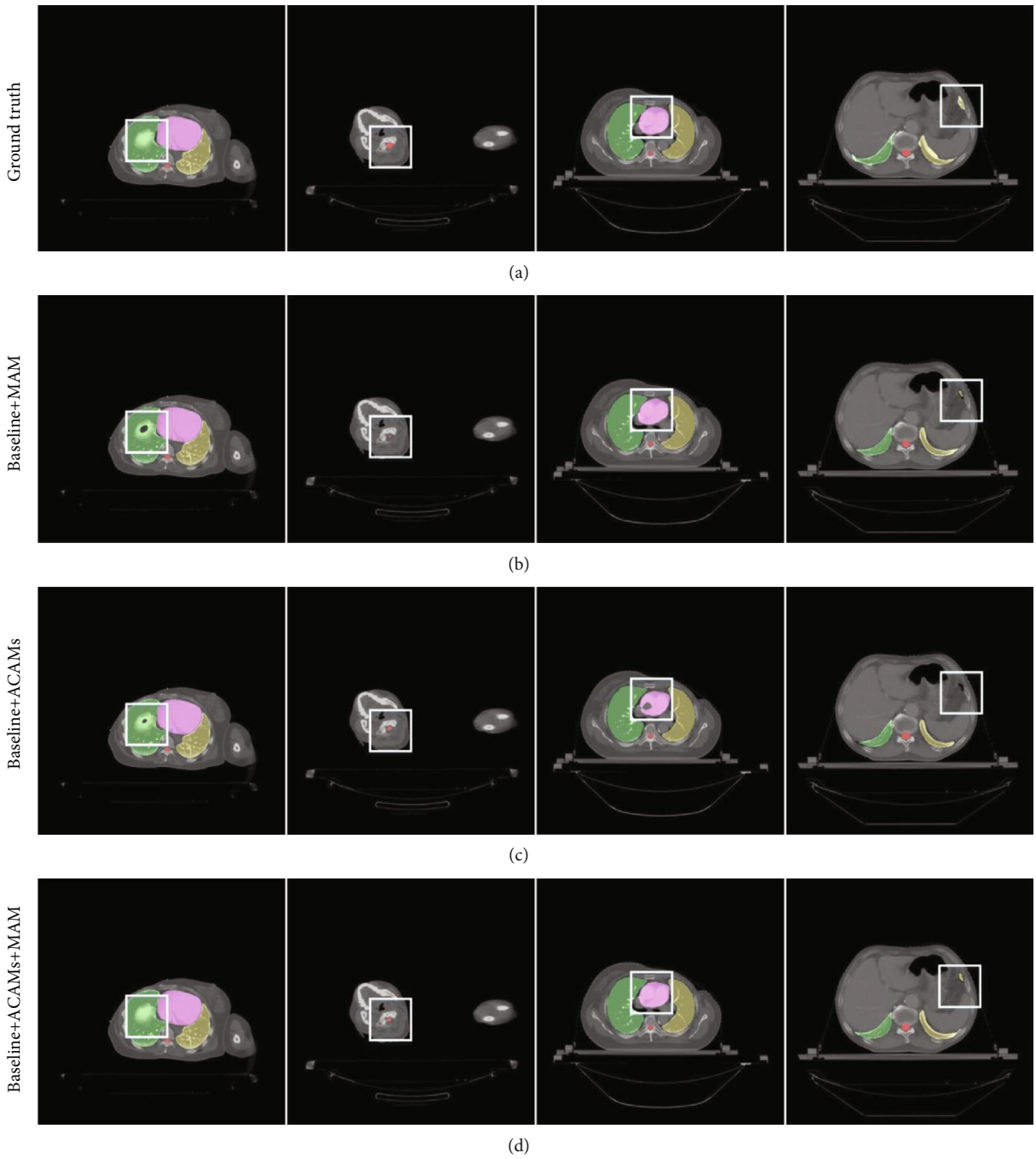


Figure 8: Qualitative results of the ablation experiment. The interesting areas on four images are framed out with white boxes.

better. A possible explanation is that more branches bring more multiscale information, which helps to capture a richer global context. Another observation is that the MAM is lightweight with negligible extra model complexity. Even if we use four branches, it will not consume much extra computation and memory resources. In all the following experiments, we use dilation of (1, 2, 4, 8) for the MAM.

3.4. Ablation Experiment. We investigate the performance of aggregating ACAMs and MAM in the proposed method. We list all the DSCs of every single organ of the in-house dataset in Table 4 and plot it in Figure 7. Aside from the conclusion in Sections 3.2 and 3.3 that both ACAMs and MAM can boost the performance of the baseline, we can get two more observations when looking into the details. First, the results obtained by jointly applying ACAMs and MAM are better than using them alone, which indicates that there is a certain complementary relationship between the two modules. Figure 8 shows qualitative results of four images, validating this observation. Second, the DSCs of small organs get more improvement, especially the spinal cord. Our method improves the DSCs of the lungs and heart by about 0.8%, while increasing the DSC of the spinal cord up to 3%.

Table 5: Comparison results of DSC (%) of the proposed method and the state-of-the-art methods on the in-house dataset.

	L-lung	R-lung	Heart	Sp-Co	Mean
U-Net [4]	96.86	96.88	94.40	89.20	94.34
PSPNet [7]	96.83	96.70	94.12	83.70	92.84
DeepLab v3+ [11]	97.25	97.01	94.10	89.07	94.36
DANet [12]	93.24	93.74	85.68	59.76	83.11
Attention U-Net [5]	97.24	97.10	94.36	91.09	94.95
Nested U-Net [6]	96.69	96.76	94.51	89.91	94.47
Ours	97.76	97.52	95.27	92.24	95.70

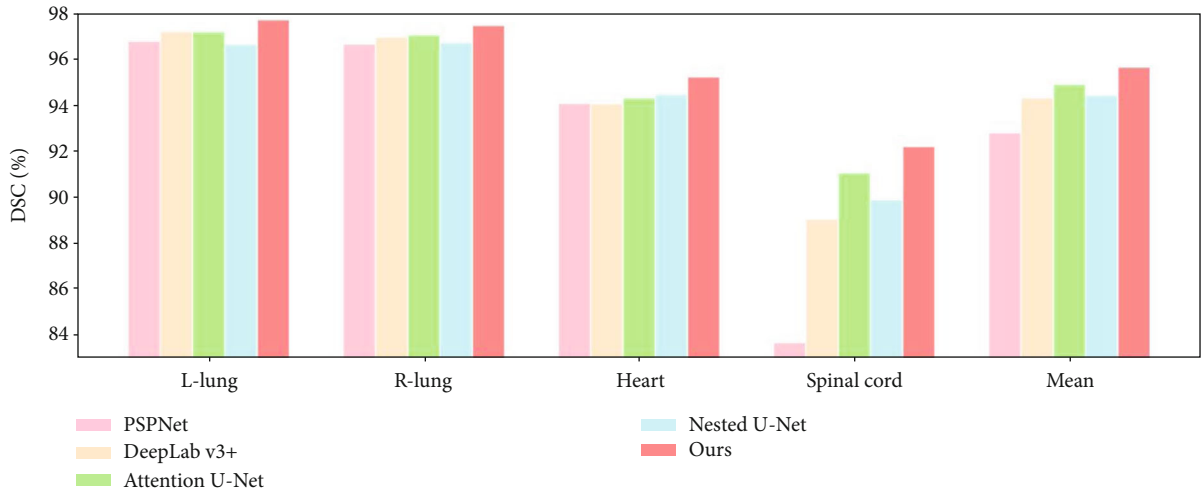


Figure 9: The DSC of each single organ.

4. Comparison Experiment

Table 5 and Figure 9 present the comparison results on the in-house dataset with several state-of-the-art methods, which are U-Net [4], PSPNet [7], DeepLab v3+ [11], DANet [12], Attention U-Net [5], and Nested U-Net [6]. The implementation of these methods can be found online (U-Net https://github.com/milesial/Pytorch-UNet, PSPNet https://github.com/hszhao/PSPNet, DeepLab v3+ https://github.com/jfzhang95/pytorch-deeplab-xception, DANet https://github.com/junfu1115/DANet, and Attention U-Net and Nested U-Net https://github.com/bigmb/Unet-Segmentation-Pytorch-Nest-of-Unets. Although DANet achieves state-of-the-art performance on the Cityscape dataset, it is not applicable to our dataset. From the table, we can see that the DSCs of the proposed method on four organs are the best among state-of-the-art methods. This can be attributed to the combination of the ACAMs and MAM, which strengthens the feature extraction capabilities. We compare our method on each individual organ with other methods, which performs the best on it. Our network additionally increases the DSC of the left lung by 0.51% over DeepLab v3+, the right lung by 0.42% over Attention U-Net, the heart by 0.76% over Nested U-Net, and the spinal cord by 1.15% over Attention U-Net.

Figure 10 shows qualitative results of four methods, which are DeepLab v3+, Attention U-Net, Nested U-Net, and our network. In the first image, the left lung looks very different from the common mirrored "C" shape. DeepLab v3+, Attention U-Net, and Nested U-Net are confused that they fail to recognize it, while our method gets a more accurate segmentation. In the second image, there is a thinner structure at the top of the left lung. Attention U-Net and Nested U-Net both ignore this structure and disrupt the continuity of the left lung during segmentation. DeepLab v3+ and our method perform well in such case. This may be due to multiscale modules applied in both methods, which is ASPP in the DeepLab v3+ and MAM in our method. In the third image, there is a noticeable small but deep depression in the left lung. Attention U-Net totally omits it and fills in this depression. DeepLab v3+ and Nested U-Net seem to observe this structure and "dig a hole" on the left lung. Our method performs the best and delineates the shape well. This further indicates that our method not only improves the segmentation of small targets but also is effective for the subtle structures. In addition, we randomly select a few cases and display the visualization results in Figure 11. In all these cases, our method has achieved satisfactory results, but there are still some defects on complex boundaries.

We observed that DANet [12] does not fit our dataset, even though it has achieved advanced results on other datasets. One possible reason is that the targets to segment in our cases have different sizes. The spinal cord is extremely small compared to other organs and background. Both the

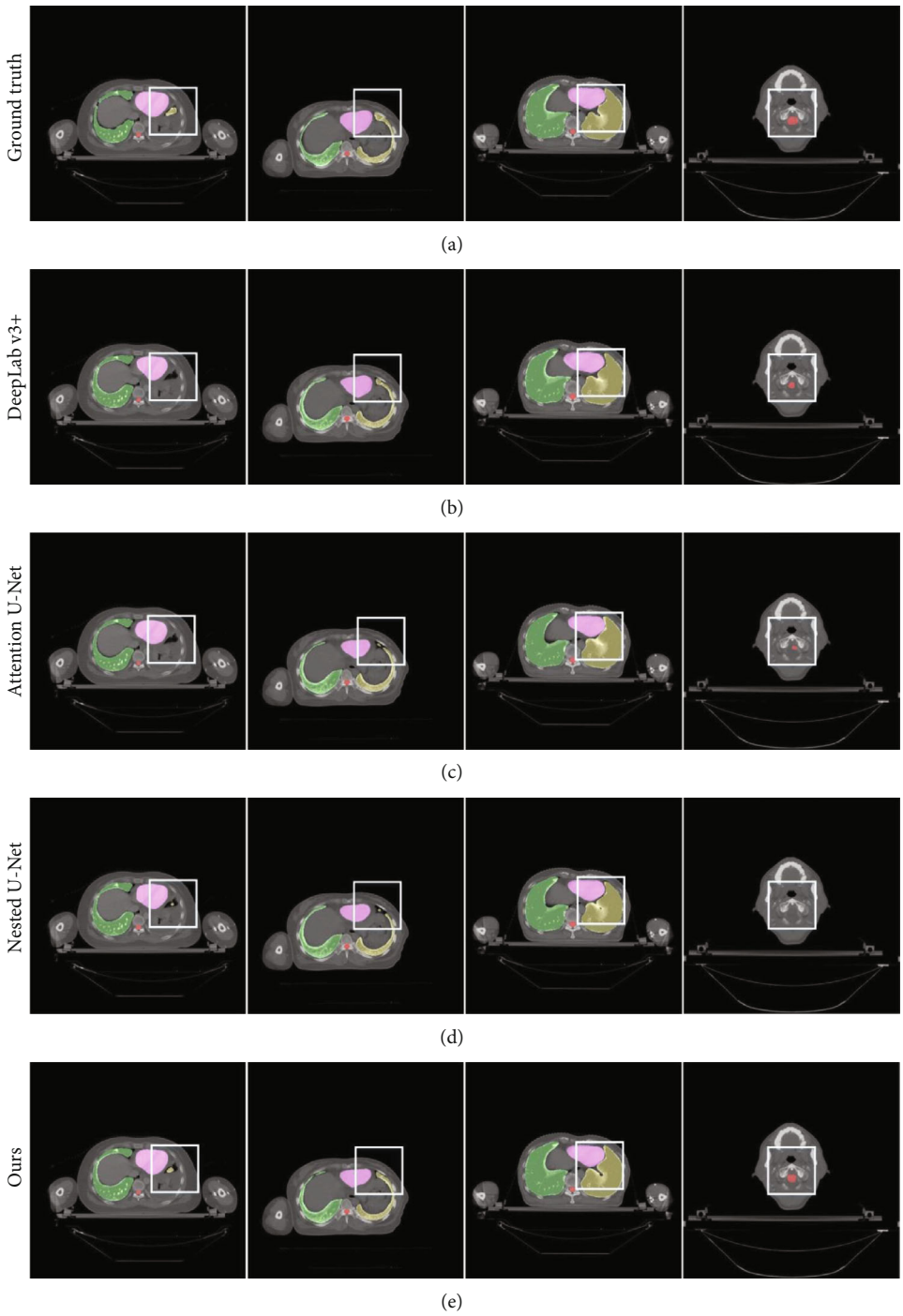


Figure 10: Qualitative results of (a) Ground Truth, (b) DeepLab v3+, (c) Attention U-Net, (d) Nested U-Net, and (e) ours.

PSPNet and DeepLab v3+ in our comparison experiment introduce pyramid pooling modules. The U-Net-based methods also introduce multiscale components due to the skip connections between different levels. However, DANet only considers the long-range dependencies between pixels but ignores the target scale information, which may become a potential factor for small object segmentation errors.

In the SegTHOR dataset, our method ranks in 4 out of 19 methods (until 6/15/2020). The esophagus, trachea, and aorta are small organs, while the heart is a relatively big organ. Except for the esophagus, our method generally achieves competitive results, which shows the effectiveness and generalization of our method. All the methods on the ranking perform the worst on the segmentation of the esophagus. We analyze the dataset and find that the esophagus is more special compared with the others, making it more challenging to segment. The esophagus is small in size and variable in shape. The boundary of the esophagus is blurred

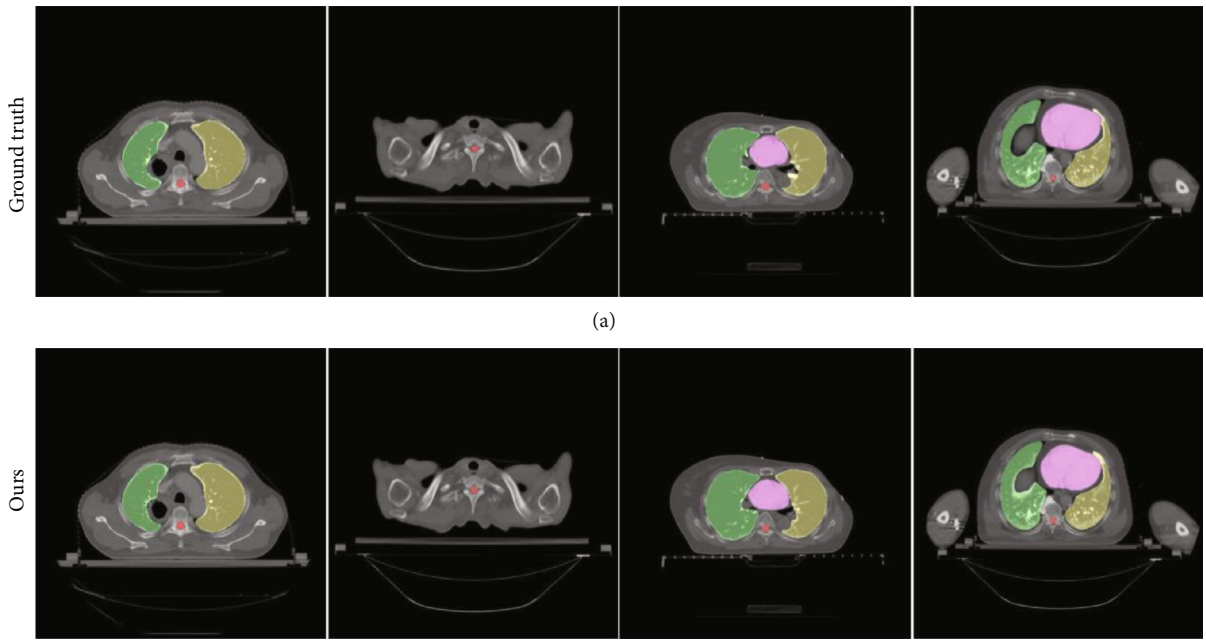


Figure 11: Visualization results of four different cases.

Table 6: FLOPs (G) and memory (10^6) are calculated on CT image with size of 512×512.

	FLOPs (G)	Memory (10^6)
U-Net [4]	123.93	1439
PSPNet [7]	262.13	2525
DANet [12]	275.42	2579
Attention U-Net [5]	266.25	2402
Nested U-Net [6]	552.01	2715
Ours	89.77	1221

due to its low contrast displayed in CT scans. Our method is not designed for such problems and is not robust enough to deal with such situations.

Model Complexity Analysis. Since the model complexity plays an important role in the design of our network, we compare the FLOPs and memory of our method with that of other state-of-the-art methods in Table 6. It is obvious that our method consumes the least computation and memory resources. The memory usage of our method is reduced nearly by 50%. The FLOPs of our method are reduced by an average of 67% except Nested U-Net. Nested U-Net has much more FLOPs due to its denser connections between convolutional layers. The reason is that the ACAMs embedded in our method compress the spatial redundancy of the feature maps, which can significantly reduce the computation cost.

5. Conclusions

In this paper, we propose an end-to-end method to segment multiple organs in CT scans. Benefiting from the compression of spatial redundancy in the applied accuracy-complexity adjustment module, the model complexity can be reduced, while achieving higher accuracy. We also present the experimental results to provide guidance for the hyperparameter selection. A nonlinear multiscale aggregation module is added after the encoder to further enrich feature representation. The combination of these two modules in the proposed method achieves higher DSC and lower complexity than several state-of-the-art methods. Such an idea provides a direction of improving performance of high-complexity models like 3D U-Net, which will be our future work.

Data Availability

The in-house dataset used to support the findings of this study was supplied by the Chinese Academy of Sciences under license and so cannot be made freely available. Requests for access to these data should be made to zhiqiangtian@xjtu.edu.cn.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by NSFC under grant No. 61876148. This work was also supported in part by the Fundamental Research Funds for the Central Universities No. XJJ2018254 and China Postdoctoral Science Foundation No. 2018M631164.

References

[1] R. Trullo, C. Petitjean, B. Dubray, and S. Ruan, "Multiorgan segmentation using distance-aware adversarial networks," *Journal of Medical Imaging*, vol. 6, no. 1, article 014001, 2019.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, MA, USA, 2015.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Munich, Germany, 2015.

[5] O. Oktay, J. Schlemper, L. Le Folgoc et al., "Attention u-net: learning where to look for the pancreas," 2018, https://arxiv.org/abs/1804.03999.

[6] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet ++: a nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, Springer, 2018.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, Honolulu, HI, USA, 2017.

[8] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, https://arxiv.org/abs/1412.7062.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[10] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, Munich, Germany, 2018.

[12] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, Long Beach, CA, USA, 2019.

[13] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 510–519, Long Beach, CA, USA, 2019.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, UT, USA, 2018.

[15] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," 2016, https://arxiv.org/abs/1611.06440.

[16] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, 2018.

[18] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, Salt Lake City, UT, USA, 2018.

[19] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, Munich, Germany, 2018.

[20] Y. Chen, H. Fan, B. Xu et al., "Drop an octave: reducing spatial redundancy in convolutional neural networks with octave convolution," 2019, https://arxiv.org/abs/1904.05049.

[21] R. Brügger, C. F. Baumgartner, and E. Konukoglu, "A partially reversible u-net for memory-efficient volumetric image segmentation," 2019, https://arxiv.org/abs/1906.06148.

[22] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, Stanford, CA, USA, 2016.

[23] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-net," in *International MICCAI Brainlesion Workshop*, pp. 234–244, Granada, Spain, 2018.

[24] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," 2015, https://arxiv.org/abs/1502.03167.

[25] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, https://arxiv.org/abs/1605.07146.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[27] Y. Huang, Y. Cheng, A. Bapna et al., "Gpipe: efficient training of giant neural networks using pipeline parallelism," in *Advances in Neural Information Processing Systems*, pp. 103–112, Curran Associates, Inc., 2019.

[28] M. Tan and Q. V. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," 2019, https://arxiv.org/abs/1905.11946.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, Santiago, Chile, 2015.

[30] W. Owadally and J. Staffurth, "Principles of cancer treatment by radiotherapy," *Surgery (Oxford)*, vol. 33, no. 3, pp. 127–130, 2015.

[31] A. L. Grosu, L. D. Sprague, and M. Molls, "Definition of target volume and organs at risk. Biological Target Volume," in *New Technologies in Radiation Oncology*, Springer, Berlin Heidelberg, 2006.

[32] S. Scoccianti, B. Detti, D. Gadda et al., "Organs at risk in the brain and their dose-constraints in adults and in children: a radiation oncologist's guide for delineation in everyday practice," *Radiotherapy and Oncology*, vol. 114, no. 2, pp. 230–238, 2015.

[33] L. D. van Harten, J. M. H. Noothout, J. J. C. Verhoeff, J. M. Wolterink, and I. I˘sgum, "Automatic segmentation of organs at risk in thoracic CT scans by combining 2D and 3D convolutional neural networks," in *Proceedings of the 2019 Challenge on Segmentation of THoracic Organs at Risk in CT Images (SegTHOR2019) co-located with the 16th International Symposium on Biomedical Imaging (ISBI)*, Venice, Italy, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 9520384, 9 pages
https://doi.org/10.1155/2020/9520384

Research Article

A Mutual Selection Mechanism of Ride-Hailing Based on Hidden Points

Yi Jiang,[1,2] **Yu Xia**,[1] **Xinyue Cheng**,[1] **and Yuntao Xu**[1]

[1]*School of Information Engineering, Yangzhou University, Yangzhou Jiangsu 225127, China*
[2]*State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

Correspondence should be addressed to Yi Jiang; jiangyi@yzu.edu.cn

Received 20 February 2020; Revised 28 May 2020; Accepted 18 June 2020; Published 15 July 2020

Academic Editor: Huimin Lu

As a new way to travel, ride-hailing has brought great convenience to the public. However, the demand expression ability of passengers in the existing ride-hailing platforms is weak, and the accuracy of the matching results is low, resulting in a large number of transaction losses and the loss of platform revenue. In this paper, we propose a kind of mutual selection mechanism of ride-hailing based on hidden points, which is composed of platform benefit calculation algorithm and driver assignment algorithm. The platform benefit calculation algorithm mainly calculates the benefits obtained by the platform after the order is completed, while the driver allocation algorithm meets the mutual selection needs of passengers and drivers. Through experiments and theoretical analysis, the mutual selection mechanism proposed in this paper has improved user satisfaction by 5%-21% compared to the traditional methods of order-snatching mechanism and order-sending mechanism, which has significantly improved passenger satisfaction and solved the problem that the existing platform cannot meet the passengers' individual travel needs and the problem that the passengers' choice is not fair. The issue of platform revenue was discussed in the context of improved passenger satisfaction.

1. Introduction

With the continuous development of modern cities, the public has higher and higher requirements for the convenience, rapidity, and comfort of transportation. In densely populated large cities, public transportation such as buses and subways sometimes cannot meet the special needs of the public. At this time, ride-hailing platforms emerge as the times require, which greatly satisfied the diverse travel needs of the public, greatly improved people's travel efficiency, and also provided more jobs. After several years of development, ride-hailing has become one of the first choices for the public to travel. Different ride-hailing platforms will adopt different matching rules, and different matching rules will lead to differences in platform matching results. The main purpose of the traditional ride-hailing platform is to improve the matching efficiency and complete more matching in a shorter time. It is reflected that the passenger side is to match the passengers in the shortest possible time. The traditional ride-hailing platform is a bipartite matching problem. In this problem, a bipartite graph $G = (U, V, E)$ can be presented, where U represents the online driver and V represents the passenger's order request. If v is assigned to u, there is an edge $e = (u, v)$. When a passenger's order request v arrives, the ride-hailing platform must make a decision to reject the request or assign v to u.

Currently, the typical ride-hailing platforms are Didi and Uber, which adopt different matching rules. The Didi platform will push the orders submitted by users to the drivers around the starting point, and the drivers will grab the orders, and the drivers who grab the orders will become the sellers of the orders. This pattern matching rule is relatively simple, and time efficiency is high, but there are some problems with this pattern: the engineer is in a state of tight time to scramble for orders, and travel safety cannot be guaranteed. In addition, malicious order grabbing and order canceling will also reduce the satisfaction of drivers and passengers, which will eventually lead to lower revenue of the platform. Unlike Didi, Uber uses a dispatch model: through a series of complex matching rules, the one who meets the needs of

customers' order sent to a matching degree is the highest driver, the driver needs to respond within a certain amount of time, and if the driver did not respond or refuse orders, the platform will order to other drivers, until a driver receives the order. Under this matching rule, Uber needs to take certain punitive measures to make drivers cautious about receiving orders and improve responsiveness. However, due to the existence of punishment measures, the number of online drivers will be smaller than that of Didi, and the ultimate matching efficiency will be lower than that of Didi.

Another problem with the existing matching rules is the problem of selection fairness. The existing matching rules mainly allow the driver to decide whether to accept or snatch the order, and the platform determines the final matching result. Passengers just issue orders without much choice. The advantage of this matching rule is that it reduces intermediate links and effectively guarantees the matching efficiency. However, the matching rule will also deprive passengers of the right to choose. Drivers can choose passengers while passengers cannot choose drivers, which is part of the reason for the recent vicious incidents of online ride-hailing. To some extent, the fairness of choice affects users' satisfaction with drivers and ride-hailing platforms.

In order to solve the problem of user satisfaction and complexity of matching rules caused by the fairness of selection, this paper proposes a mutual selection mechanism based on hidden points. This paper is divided into six chapters: the second section introduces the related work. The third section introduces the model and the details of mechanism. The fourth section introduces the algorithm and analysis of the mechanism. The fifth section introduces the experiment part. The final section summarizes the full text. The introduction should be succinct, with no subheadings. Limited figures may be included only if they are truly introductory and contain no new results.

2. Related Work

Compared with the traditional way of travel, online ride-hailing has certain advantages. Chen et al. [1]proposed a heuristic algorithm, which showed that ride-sharing is an effective method to reduce the number of vehicles required for travel and vehicle mileage. In particular, when the degree of participation is high and the starting point and destination of the journey are more spatially concentrated, carpooling can reduce more vehicle miles traveled and the number of vehicles used. Ride-hailing platforms also need to have a strategy to attract potential users, so that the operation of the platform enters a virtuous cycle. Wang et al. [2] studied the factors influencing potential users to use online ride-hailing platforms. Henao and Marshall [3] analyzed the relevant data of drivers and concluded that after completing the order, drivers should stop and wait for the next ride request instead of driving to the active area, unless doing so can reduce the ride request waiting time by at least 30%. Based on the SOR model, Yang and Chen [4] constructed a factor model of copassenger satisfaction and analyzed customer satisfaction from four dimensions: perceived usefulness, perceived ease of use, perceived travel risk, and perceived service quality. Gilibert et al. [5] contributed to the new DRT mechanism by identifying user needs and market opportunities. Segal-Halevi et al. [6] proposed a real multiunit bilateral auction mechanism, in which, by dividing the market into left and right submarkets, eventually, the mechanism are prior-free, dominant-strategy incentive-compatible, individually rational, and budget-balanced. Lee et al. [7] proposed a taxi dispatching system based on real-time traffic conditions. Miao et al. [8] combined data information with real-time control decision to balance the minimum total free driving distance between an idle taxi and the minimum free driving distance. By analyzing Uber's system architecture, Watanabe et al. [9] found the importance of price decline and order increase in a virtuous cycle to the platform architecture. Xu et al. [10] proposed an order scheduling algorithm based on a large-scale car-hailing platform, aiming to provide a more effective method for optimizing resource utilization and user experience from a global and longer-term perspective and significantly improve the allocation efficiency of the platform. Henao and Marsall [11] analyzed the impact of ride-hailing on traffic efficiency and concluded that the vehicle mileage brought by ride-hailing was much higher than that without ride-hailing. Pham et al. [12] enable privacy-conscious riders to achieve levels of privacy that are not possible in the current RHSs and even in some conventional taxi services, thereby offering a potential business differentiator. Young and Farber [13] show that ride-hailing is too minute and inconsequential to influence the ridership level of other more substantial modes of travel overall, when considering specific market segments; the rise of ride-hailing corresponds to a significant decrease in taxi ridership and a rise in active modes of travel. And it will have a much more pronounced effect on the level of ridership of other modes as well. de Souza Silva et al. [14] show that the majority of ridesourcing trips is replacing taxi and public transport trips. Safety and cost are the main reasons that influence the decision of sharing trips via ride-splitting. The use of larger vehicles for sharing trips can introduce competition with the public transport systems. The ridesourcing interference on collective public transportation may be more noticeable than on individual public transport (taxis), given the much greater demand for the former. Mäntymäki et al. [15] say the stark power disparity between workers and the platform is, in turn, a major source of discontent among workers and put forward two key dimensions of work relations in the context of platform-enabled work: digital temporality and algorithmic administration. Guo et al. [16] assess the impact of ride-hailing platforms' market entry on new car purchases in the presence of platform competition and found the two competing platforms may have provided subsidies to drivers such that more people purchased new cars in order to sign up as drivers. Ma et al. [17] report on the development of an integrated model to investigate how perceptions of risk play into a person's decision to stop using a particular ride-hailing service and users' trust in drivers has a positive effect on users' trust in the ride service platform and their attitude towards the platform; users' trust in the platform positively affects their attitude towards the platform.

In general, most of the relevant studies focus on platforms and drivers, with demand forecasting and vehicle scheduling as the focus [18–21]. However, few studies have addressed the problem of passenger satisfaction from the perspective of passengers. The passenger satisfaction largely determines the retention rate of users. If the problem of user satisfaction cannot be solved well, the platform will lose users and eventually lead to lower revenue. Therefore, this paper focuses on the research of passenger satisfaction and proposes a mutual selection mechanism for ride-hailing based on hidden points. This mechanism improves the effectiveness of the platform on the basis of solving passenger satisfaction.

3. Model and Mechanism Design

3.1. Model. There are three main roles in the ride-hailing matching mechanism: driver (u), passenger (v), and ride-hailing platform. A driver-owned vehicle can be seen as a commodity that can take passengers to their destination. As the provider of services, the driver can act as a seller in the market. Passengers issue demands and provide money. As the party buying services, passengers issue service requests to ride-hailing platforms and provide fees after the driver sends them to their destination, which can be understood as the buyer in the market. The ride-hailing platform provides a price function, matches drivers and passengers, and charges a certain handling fee as revenue. Both drivers and passengers are users of the platform. As the party providing the platform, the ride-hailing platform can be understood as a market maker in the market (see Figure 1).

Each user i in the matching mechanism has its own value function v_i, which is initialized to $v_i(0) = 0$. The user's value function is determined by the benefits brought by the trip. The value function of the driver is the cost and profit composition of the service. In this paper, it is assumed that the platform has n orders, the price function set by the platform is p, and the travel distance is m in an order $j(j \in [0, n])$. So for buyer v, $\text{Gain}_{j,v}(m, p) = v_v(m) - p(m)$. For seller u, $\text{Gain}_{j,u}(m, p) = p(m) - v_u(m) - h$. h represents the no-load cost of the driver. This distance is only cost and no benefit.

The matching mechanism generates a matching result $G = (U, V, E)$ and a transaction price function p according to the passenger's demand order. In the matching mechanism proposed in this paper, the matching result is generated by the mutual selection of the driver and the passenger. Both of them need to accept the matching result and be responsible for the matching result and pay a certain transaction fee to the platform f_j, where $f_j = f_u + f_v$. If a matching mechanism is budget balanced, then $\sum_{j \in \text{Order}} f_j \geq 0$. If each user receives a positive return, then a mechanism is personally rational; that is, $\text{Gain}_j(m, p) - f_j \geq 0$ is inevitable. If a user cannot increase his own income by falsely reporting the valuation, then, this mechanism is real. This paper defines the passenger satisfaction function as s_j. User satisfaction will bring some hidden benefits to the platform. Then $\text{Utility} = \sum_{j=1}^{n} f_j + s_j - C$, where C represents platform operating costs.

3.2. Mechanism Details. Under this mechanism, passengers send a service request order to the online ride-hailing platform at time t to request service. Based on the principle of proximity, the online ride-hailing platform first uses the hidden point matching mechanism to first select the X drivers closest to the user's departure point, then filter out m drivers close to the user's hidden points, order will be pushed to the m drivers, n drivers out of m drivers ($n \leq m$), then the platform pushes the information of the n drivers (gender, distance, vehicle condition, vehicle type, driving speed, etc.) to the user, and the user selects the service driver. At this time, the driver and passenger are successfully matched, that is, the value $e = (u, v)$.The driver is a reusable resource [22]. The driver grabs orders at time t', and the passenger decides to serve the driver at time t''.The order is completed at time t''', then the driver will rejoin the matching queue after time t''', and the passenger will occupy the time $t''' - t$, which is the service time. The occupancy time of driver u matching with passenger v depends on the user type (such as destination) of v and the time when the matching occurs (peak time may be significantly different from off-peak time).

Hidden points' mechanism using weighted algorithm calculates hidden points for driver and passenger separately. Hidden points for drivers include the order grabbing rate, order completion rate, user evaluation, and complaint rate. The order grabbing rate refers to the driver's acceptance rate of orders pushed by the platform. Order completion rate refers to the proportion of the number of orders grabbed by the driver and completed by the driver in the number of orders grabbed by the driver. The order grabbing rate and order completion rate will combine to affect the hidden score of the driver. User evaluation includes multiple scores, including passenger satisfaction with service, driver's service attitude, vehicle environment, vehicle speed, and whether to arrive on time, and a comprehensive evaluation score is obtained through a weighting algorithm. Both the driver and the platform have the right to view user evaluations. Drivers can improve their service quality based on user evaluations, provide better services, improve their hidden points, and obtain higher profits. The complained rate refers to the proportion of the driver's complained orders in the number of orders he completed. The complaint rate is proportional to the hidden ratio. The passenger's complaint against the driver will be handled manually by the online car-hailing platform. The passenger's evaluation of the platform will be reflected in the user satisfaction later. Passenger hidden score includes complaint rate, usage rate, and driver evaluation. Among them, the complaint rate refers to the proportion of passenger complaint orders in all the submitted orders. When users encounter some problems that are difficult to coordinate with the driver, they can use complaints to protect their rights and interests. Under normal circumstances, the passenger's complaint rate will be maintained at a normal level, but if the passenger's complaint rate exceeds the normal level, it will be classified as a poor-quality passenger, and the complaint rate is inversely proportional to the hidden share. Utilization rate refers to the frequency with which passengers use the ride-hailing platform, and the utilization rate

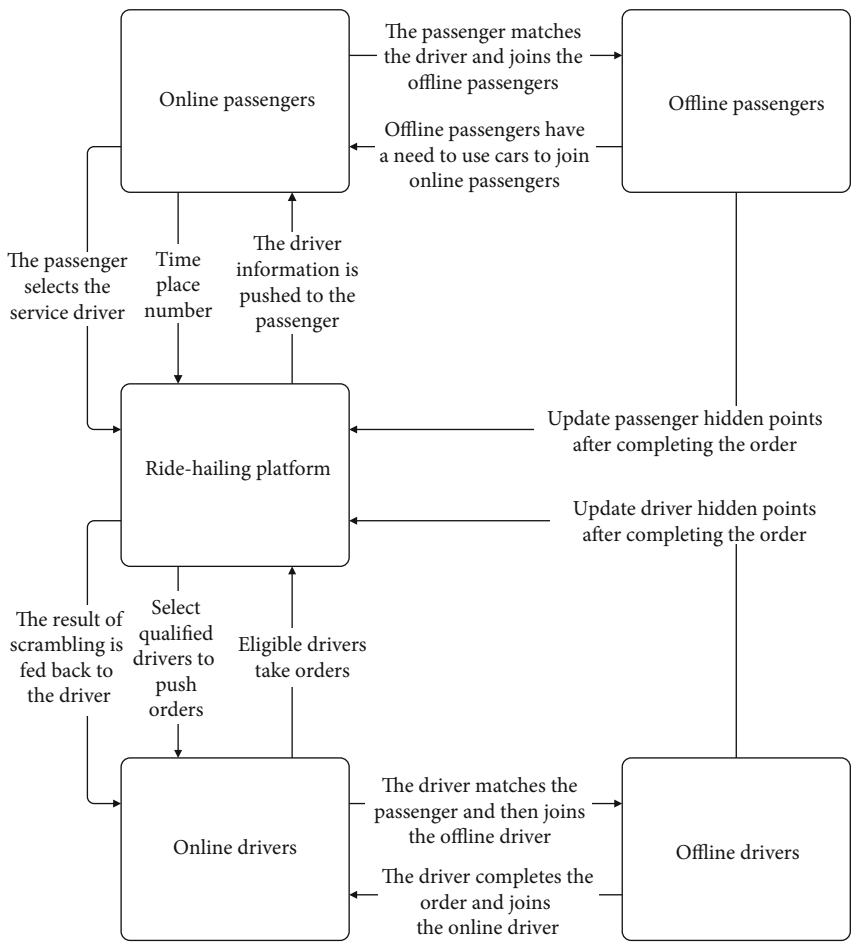


Figure 1: Structure diagram of ride-hailing matching.

evaluation is proportional to the concealment. Driver evaluation refers to the driver's evaluation of passengers. The higher the evaluation, the higher the hidden score. Drivers and passengers evaluate each other. The evaluation is not displayed in real time (that is, it can be evaluated after the order is completed), but it can be evaluated one minute after the order is completed, and the evaluation content is not displayed in real time, but only after the mutual evaluation of both parties is seen. This reduces false ratings to a certain extent, because under the scoring mechanism of the existing platform, the order can be evaluated once the order is completed, and some drivers will require passengers to give a full score evaluation, even if some passengers are not satisfied with the itinerary, only a full score evaluation. Delaying the evaluation time can to a certain extent ensure that passengers leave the driver's visible range for evaluation, which improves the authenticity of the evaluation.

(i) Hidden points for drivers:

$$\mathrm{HP}_{d_i} = \sum_{k\in[K]} \sum_{n=1,2} k_i^n w_{k^n} \frac{1}{k_i^3} k_i^4, \tag{1}$$

where for drivers and K is the order scrambling rate, order receiving completion rate, complained rate, and user evaluation.

(ii) Hidden points for passengers:

$$\mathrm{HP}_{p_i} = \mathrm{k}_i^1 \frac{1}{k_i^2} k_i^3, \tag{2}$$

where for passengers, K is the usage rate, complaint rate, and driver evaluation.

In this paper, assuming the driver's Supply for Supply_u, and the Demand of users for Demand_v, generally speaking, we have excess demand ($\mathrm{Demand}_v > \mathrm{Supply}_u$), excess supply ($\mathrm{Demand}_v < \mathrm{Supply}_u$), and Supply and Demand which are equal ($\mathrm{Demand}_v = \mathrm{Supply}_u$). In the ride-hailing platform, when $\mathrm{Demand}_v > \mathrm{Supply}_u$, it means that it is the current time to peak, and when $\mathrm{Demand}_v < \mathrm{Supply}_u$, it means that the current session is a low period. In this paper, we only discuss the case of oversupply, where the number of drivers far outweighs the number of user orders.

The step is shown Figure 2.

Step 1. Passengers send a service request to the platform, platform passenger departure, arrival, classification of hidden information to be obtained, and then on passenger source, screening from the starting point of the recent X driver, according to the X driver hidden points; filter out m drivers

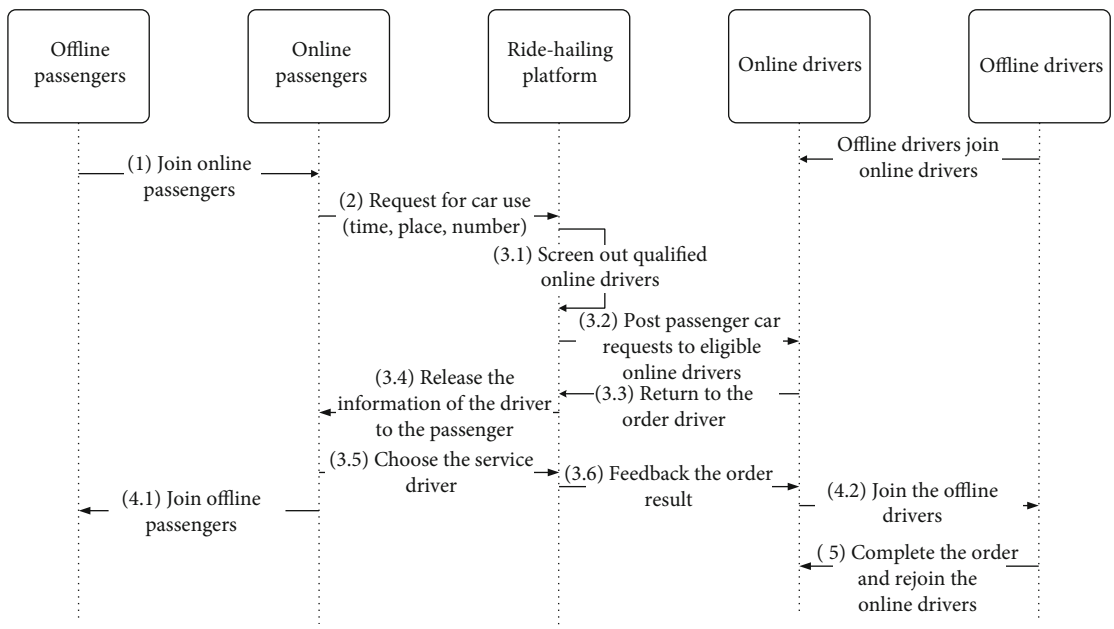


Figure 2: Timing sequence diagram of ride-hailing matching.

who are close to the user's hidden points; and then push user orders to these m drivers.

Step 2. Among the m drivers, there will be n drivers scrambling for orders ($n \leq m$), and then the platform will push the information of these n drivers (gender, distance, vehicle condition, vehicle type, average driving speed, etc.) to the user, who will finally choose the service driver. After the passenger selects the service driver, the platform matches the two, adds the service driver to the offline pool, returns the information of the successful order snatcher to the driver, and returns the information to the $n-1$ order snatcher to inform them of the failed order snatcher.

Step 3. After the journey is completed, the platform calculates the fee that passengers should pay to the driver and the transaction fee f_j that users need to pay to the platform based on the unit price p of the mileage. The transaction fee f_j is paid by the passenger and the driver, respectively. After the transaction is completed, the driver and passenger evaluate each other about the service, and the evaluation results are hidden from both parties on the day and visible to both parties on the following day.

3.3. Algorithm and Analysis. This section introduces a matching algorithm of driver-passenger mutual selection based on the hidden points, which is aimed at designing a matching rule. On the basis of ensuring the matching efficiency, the algorithm enables passengers to have certain right of demand expression, improves user satisfaction, and finally increases the revenue of the platform.

For each driver, there is a certain probability to decide whether to snatch the order. In this paper, the probability of the driver's decision to snatch the order is defined as x_e. For each passenger, each driver has a matching weight value; that is, matching $e = (u, v)$ is related to a weight w_e. Since the driver is a reusable resource, let $t' < t$. In formula (4), the first term represents that u is unavailable at time t, and the second term represents that u was assigned to other passengers at time t. The probability that is less than u is available at time t.

$$\begin{aligned} \max \quad & \sum_{t\in[T]}\sum_{e\in E} x_e w_e Uti \\ \text{subject to} \quad & 0 \leq x_e \leq 1 \forall e \in \mathrm{E}, \quad t \in [T], \end{aligned} \tag{3}$$

$$\sum_{t'<t}\sum_{e\in E_u} x_{e,t'} + \sum_{e\in E_u} x_{e,t} \leq 1 \forall u \in U, \quad t \in [T]. \tag{4}$$

The distribution benefit algorithm of the ride-hailing platform is shown in Algorithm 1, where Θ_t represents the set of drivers in the driver pool waiting at time t. $U = \{u, g_u, S_u\}$ is the driver information set, where u represents the number of the driver, g_u represents the hidden score of the driver numbered with u, S_u represents the state of the driver numbered with u, $S_u = 1$ means the driver is online and waiting for orders, and $S_u = 0$ means the driver is offline or carrying passengers. $\Omega_j\{l_v, \mathrm{t}, g_v, S_j\}_{i\in I}$ said orders j collection of information, l_v said user v origin, t time, g_v said user v hidden points, and S_j said orders j order status. The Utility $= \sum_{j=1}^{n} f_j + s_j - C$ represents the platform's revenue.

When passengers v order submission requirements for platform, system at time $t \in [a_j, d_j]$ in the distribution of the driver to the user. The system allocates drivers through an allocation algorithm, as shown by Algorithm 2. After the order is completed, the platform needs to modify the status of the driver and update the hidden point information of the driver pool and calculate the revenue of the order.

Input: θ_t ; $U = \{u, g_u, S_u\}$; $\Omega = \{l_v t, g_v, S_i\}_{i \in I}$; P ; T
Output: θ_t ; $U = \{u, g_u, S_u\}$; $V = \{v, g_v\}$; $Utility$
$\theta_t = \{1 \cdots M\}, \forall m \in M, \forall t \in [1, T]$
while *Receiving an order* θ_t **do**
  for $t \in [a_j d_j]$ **do**
    Execute Allocation Algorithm shown in Algorithm 2
    while $S_1 = 1$ **do**
      $S_u = 1$
      Update the θ_t, g_u, g_v
      $Utility = f_j + s_j - c$
    end
  end
end

Algorithm 1: Online platform.

Input: θ_t ; k ; $U = \{u, g_u, S_u\}$
Output: e = (u, v)
$\bar{d} = (d_{t\ \min} + d_{\min})/2$
$\theta_t = \theta_t \setminus$ (the *distance that satisfying* $\bar{d} > k$)
$U_1 = \theta_t \setminus (|g_u - g_v| > x)$
If $U_1 \neq \theta$ **then**
  Send order information to the driver in U_1
  Driver grab the order and get a new optional collection $U_2 = \{u, g_u, S_u\}$
  while $U_2 \neq \theta$ **do**
    Send this drives' information to user to select
    if $t \in [a_j d_j]$ and $S_u = 1$ **then**
      $S_u = 0$
      $e = (u, v)$
      return
    else
      Exclude the driver satisfying $S_u = 0$ from U_2
      T=0
    end
  end
else
  return
end

Algorithm 2: Allocation algorithm.

In Algorithm 2, k represents the maximum screening distance between the driver and the passenger defined by the system during the primary election. According to the source information of users in the order, the system screened out all drivers whose distance from the source is less than k. Distance $\bar{d}$ was driver's fastest time average distance and the shortest distance. $e = (u, v)$ represents the successful pairing of the driver and passenger. x is the score difference set by the system. In order to provide better service, the system will automatically match the score difference between passengers and drivers within x. There may be multiple passengers using ride-hailing services at the same time, and a driver may also meet the needs of multiple passengers. The allocation algorithm sets a timer, and the user must select the driver within the valid time to prevent the driver from being selected by other passengers or getting pulled off the line abnormally. Because of the double-option mechanism, the driver is likely to have been selected by user B before being selected by user A, which also encourages the user to choose as soon as possible. If the driver has been selected, the system pushes the set of drivers back to the user until the match is successful or there is no driver to choose from in the driver pool.

Theorem 1. *The matching mechanism is prior-free, dominant-strategy, incentive-compatible, individually rational, and budget-balanced.*

Proof. A priori free is determined by the mechanism, which does not need to collect users' value information. In dominant-strategy and incentive-compatible, for passengers, only by sending an order can they get the service. The behavior of sending an order is the dominant strategy, while for drivers, no matter how other drivers choose, the behavior of scrambling for an order is the dominant strategy. Under

this mechanism, passengers and drivers are unable to increase their earnings by inflating the valuation. The price is set by the platform, and neither side can influence the pricing. Therefore, the mechanism is incentive compatible. Through the platform, passengers and drivers cooperate with each other and finally realize the maximization of their respective interests, which is obviously rational for individually rational. The mechanism is budget balanced; that is, the platform will not lose money and can obtain transaction fees from the transaction.

Theorem 2. *Under this algorithm, the improvement of passenger satisfaction will bring more hidden benefits to the platform.*

Proof. Under this mechanism, passengers' satisfaction will be enhanced due to their breeding needs being satisfied to a certain extent, and their own needs are attached importance to the platform, and the fairness of platform matching is guaranteed. The passengers' satisfaction will drive new passenger to join the platform, and the increasing of the number of passengers will make more drivers choose to use the platform, and an increase in the number of driver will attract more people to join platform into the benign circulation state, eventually rise in the number of users, and bring more benefits to the platform.

User satisfaction consists of three parts: matching time, matching efficiency, and evaluation,

$$s_j = \sum_{k\in[K]} \sum_{n=1}^{3} L \frac{1}{\left(t'' - t\right)} x_e w_e k^n w_{k^n}, \tag{5}$$

where L is a constant and K is the platform evaluation, service quality, and demand satisfaction.

3.4. Experiments. In order to verify the effectiveness of the mechanism proposed herein, this paper simulates a scenario in which there is a ride-hailing platform with 1 passenger and 20 drivers randomly distributed around passengers. The experiment simulates three different matching methods (platform A adopted the pattern of scrambling for orders, platform B adopted the pattern of dispatching orders, and platform C adopted the matching rules proposed in this paper) and finally generates 3 sets of data, each set of data with 50 records. The records generated by different matching rules are different. The simulation data of platform A contains the matching time of the order, driver's order grabbing rate, user evaluation, and other data; the data of platform B contains the order matching time, the driver's order acceptance rate, and the user evaluation and other data; the data simulated by platform C includes the matching time of orders, driver's rate of obtaining orders, user's consent rate, user's evaluation of platform, user's evaluation of itinerary, and user's demand satisfaction. The simulated data includes the matching time, matching efficiency, users' evaluation of the platform, users' evaluation of the journey, and users' satisfaction degree for each order under different rules.

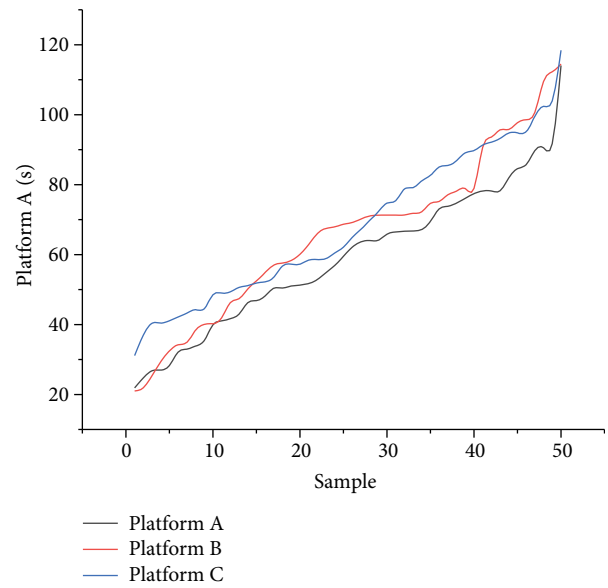


Figure 3: Comparison of platform matching time.

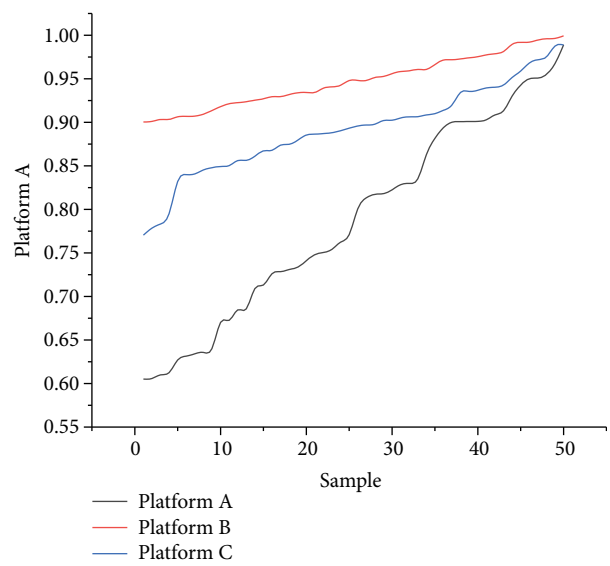


Figure 4: Comparison of platform matching rate.

This experiment assumes that under three platforms, the number of drivers is much larger than that of passengers, which ensures that the order matching efficiency under different platforms will not be affected by the relationship between supply and demand. Three different matching rules were used to calculate the satisfaction of each group of 50 data under each matching rule, and the experimental results were as follows.

From the above figure, we can see that the overall matching time of platform C is longer than those of platform A and platform B. The average matching time of platform C is 15% longer than that of platform A and 3.6% longer than that of platform B (see Figure 3). The overall matching rate of platform C is better than that of platform A and inferior to that of platform B. The average matching rate of platform C is 6% lower than that of platform B and 12% higher than that of platform A (see Figure 4). The overall evaluation of

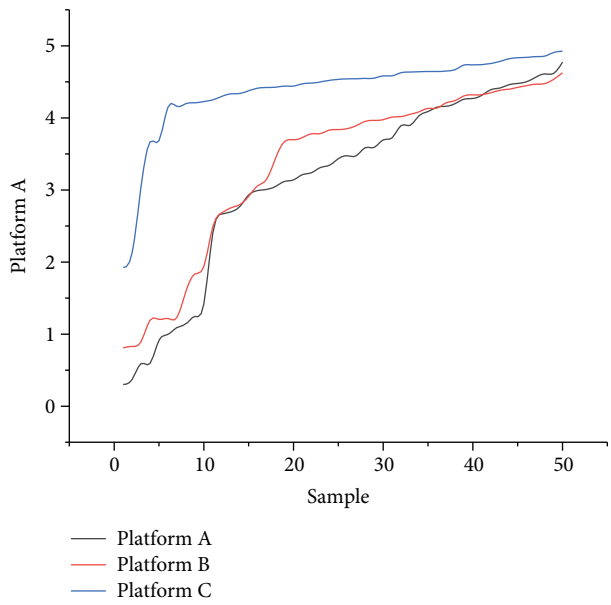


Figure 5: Comparison of platform evaluation.

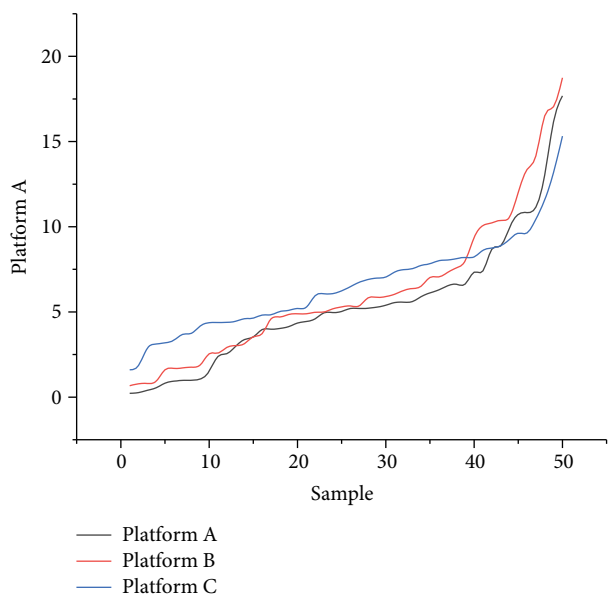


Figure 6: Comparison of platform satisfaction.

platform C is better than those of platform A and platform B. The average evaluation of platform C is 39% higher than that of platform A and 30% higher than that of platform B (see Figure 5). The overall user satisfaction of platform C is better than those of platform A and platform B. The satisfaction of platform C is 21% higher than that of platform A and 5% higher than that of platform B (see Figure 6).

Since platform A adopts the order scrambling rule, the platform will send orders to 20 drivers, and 20 drivers will scramble the order. The competition ratio is 1/20. Although the matching speed of platform A has a great advantage over those of platforms B and C, the final matching result is unpredictable and cannot reflect the expression of demand by passengers. The driver's competition is fierce and passenger satisfaction is lower than that of platform B and C levels. Since the immediate needs of passengers cannot be met, the loss of passengers will be caused, and the loss of passengers will further lead to the loss of drivers, resulting in a vicious cycle, and the platform's final revenue will be reduced.

Platform B adopts the dispatch rules; then the platform will use its complex algorithm to select one driver from 20 drivers to send an order with a competition ratio of 0. The final matching result is determined by the algorithm itself and has a certain degree of controllability. The matching result is more in line with user expectations than platform A, but it is difficult to meet the immediate needs of passengers. Its user satisfaction is higher than that of platform A, and the matching time is no different than that of platform C. But user satisfaction is not as good as that of platform C.

Platform C, which adopts the mechanism proposed in this article, will choose to send orders to 10 drivers from 20 drivers. The final competition ratio is determined by the number of single drivers but must be less than 1/10. Driver's competition ratio is better than that of platform A and inferior to that of platform B. However, users can choose the driver that best meets their needs among the n drivers who grab the order. The user's immediate needs can be met to a certain extent, and user satisfaction will rise. The increase in user satisfaction will bring more new passengers to the platform, and the addition of new passengers will also bring more drivers. Ultimately, the platform's revenue will gradually increase as user satisfaction increases.

4. Conclusion and Future Directions

Inability to guarantee the retention rate of users and the amount of new users introduced are the problems that the ride-hailing platform needs to solve urgently. The key to solving this problem is to improve users' satisfaction and make the platform enter a virtuous cycle, thus increasing platform revenue. In order to solve this problem, this paper proposes an allocation mechanism. Through comparative experiments, it is proved that this mechanism can improve user satisfaction, make the platform enter a virtuous circle, obtain more benefits, and solve the problem of fairness selection under the existing matching rules.

However, this paper focuses on passenger satisfaction, and the users of the platform include not only passengers but also drivers. How to improve driver satisfaction will be our next step. At the same time, although the price function is introduced in this paper, it is not discussed further. How to develop a reasonable price function will also be the focus of the next research work.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61872313, in part by the Key Research Projects in Education Informatization in Jiangsu Province under Grant 20180012, in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant KYCX18\2366, in part by the Yangzhou Science and Technology under Grant YZ2018209 and Grant YZ2019133, in part by the Yangzhou University Jiangdu High-End Equipment Engineering Technology Research Institute Open Project under Grant YDJD201707, and in part by the Open Project in the State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, under Grant 1907.

References

[1] W. Chen, M. Mes, M. Schutten, and J. Quint, "A ride-sharing problem with meeting points and return restrictions," *Transportation Science*, vol. 53, no. 2, pp. 401–426, 2019.

[2] Y. Wang, J. Gu, S. Wang, and J. Wang, "Understanding consumers' willingness to use ride-sharing services: the roles of perceived value and perceived risk," *Transportation Research Part C: Emerging Technologies*, vol. 105, pp. 504–519, 2019.

[3] A. Henao and W. E. Marshall, "An analysis of the individual economics of ride-hailing drivers," *Transportation Research Part A: Policy and Practice*, vol. 130, pp. 440–451, 2019.

[4] J.-C. Yang and D.-D. Chen, "Influencing factors of customer satisfaction towards ride-sharing," *Ecological Economy*, vol. 15, no. 2, pp. 88–94, 2019.

[5] M. Gilibert, I. Ribas, N. Maslekar, C. Rosen, and A. Siebeneich, "Mapping of service deployment use cases and user requirements for an on-demand shared ride-hailing service: Moia test service case study," *Case Studies on Transport Policy*, vol. 7, no. 3, pp. 598–606, 2019.

[6] E. Segal-Halevi, A. Hassidim, and Y. Aumann, "Muda: a truthful multi-unit double-auction mechanism," in *Thirty-Second AAAI Conference on Artifcial Intelligence*, pp. 1193–1201, AAAI Press, New Orleans, 2018.

[7] D.-H. Lee, H. Wang, R. L. Cheu, and S. H. Teo, "Taxi dispatch system based on current demands and real-time traffic conditions," *Transportation Research Record*, vol. 1882, no. 1, pp. 193–200, 2004.

[8] F. Miao, S. Han, S. Lin et al., "Data-driven robust taxi dispatch under demand uncertainties," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 1, pp. 175–191, 2019.

[9] C. Watanabe, K. Naveed, and P. Neittaanmäki, "Co-evolution of three mega-trends nurtures un-captured GDP – Uber's ride- sharing revolution," *Technology in Society*, vol. 46, pp. 164–185, 2016.

[10] Z. Xu, Z. Li, Q. Guan et al., "Large-scale order dispatch in on-demand ride-hailing platforms: a learning and planning approach," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 905–913, New York, 2018.

[11] A. Henao and W. E. Marshall, "The impact of ride-hailing on vehicle miles traveled," *Transportation*, vol. 46, no. 6, pp. 2173–2194, 2019.

[12] A. Pham, I. Dacosta, B. Jacot-Guillarmod et al., "PrivateRide: a privacy-enhanced ride-hailing service," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 2, pp. 38–56, 2017.

[13] M. Young and S. Farber, "The who, why, and when of uber and other ride-hailing trips: an examination of a large sample household travel survey," *Transportation Research Part A: Policy and Practice*, vol. 119, pp. 383–392, 2019.

[14] L. A. de Souza Silva, M. O. de Andrade, and M. L. Alves Maia, "How does the ride-hailing systems demand affect individual transport regulation?," *Research in Transportation Economics*, vol. 69, pp. 600–606, 2018.

[15] M. Mäntymäki, A. Baiyere, and A. K. M. N. Islam, "Digital platforms and the changing nature of physical work: Insights from ride-hailing," *International Journal of Information Management*, vol. 49, pp. 452–460, 2019.

[16] Y. Guo, X. Li, and X. Zeng, "Platform competition in the sharing economy: understanding how ride-hailing services influence new car purchases," *Journal of Management Information Systems*, vol. 36, no. 4, pp. 1043–1070, 2019.

[17] L. Ma, X. Zhang, X. Ding, and G. Wang, "Risk perception and intention to discontinue use of ride-hailing services in China: taking the example of DiDi Chuxing," *Transportation Research Part F: Psychology and Behaviour*, vol. 66, pp. 459–470, 2019.

[18] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: a fast and lightweight network for single-image super resolution," *IEEE Transactions on Cybernetics*, pp. 1–11, 2020.

[19] H. Lu, Y. Li, S. Mu, D. Wang, H. Kim, and S. Serikawa, "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2315–2322, 2018.

[20] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain intelligence: go beyond artificial intelligence," *Mobile Networks and Applications*, vol. 23, no. 2, pp. 368–375, 2018.

[21] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2400–2413, 2020.

[22] J. P. Dickerson, K. A. Sankararaman, A. Srinivasan, and P. Xu, "Allocation problems in ride-sharing platforms: online matching with ofine reusable resources," in *Thirty-Second AAAI Conference on Artifcial Intelligence*, pp. 1007–1014, AAAI Press, New Orleans, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8895429, 15 pages
https://doi.org/10.1155/2020/8895429

Research Article

Leveraging Deep Learning Techniques for Malaria Parasite Detection Using Mobile Application

Mehedi Masud,[1] Hesham Alhumyani,[1] Sultan S. Alshamrani,[1] Omar Cheikhrouhou,[1] Saleh Ibrahim,[2,3] Ghulam Muhammad,[4] M. Shamim Hossain,[5] and Mohammad Shorfuzzaman[1]

[1]*College of Computers and Information Technology, Taif University, Taif 21974, Saudi Arabia*
[2]*Electrical Engineering Department, Taif University, Saudi Arabia*
[3]*Computer Engineering Department, Cairo University, Egypt*
[4]*Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia*
[5]*Department of Software Engineering, College of Computer and Information Sciences, King Saud University, King Saud University, Riyadh 11543, Saudi Arabia*

Correspondence should be addressed to M. Shamim Hossain; mshossain@ksu.edu.sa

Received 26 March 2020; Revised 19 May 2020; Accepted 2 June 2020; Published 8 July 2020

Academic Editor: Yin Zhang

Malaria is a contagious disease that affects millions of lives every year. Traditional diagnosis of malaria in laboratory requires an experienced person and careful inspection to discriminate healthy and infected red blood cells (RBCs). It is also very time-consuming and may produce inaccurate reports due to human errors. Cognitive computing and deep learning algorithms simulate human intelligence to make better human decisions in applications like sentiment analysis, speech recognition, face detection, disease detection, and prediction. Due to the advancement of cognitive computing and machine learning techniques, they are now widely used to detect and predict early disease symptoms in healthcare field. With the early prediction results, healthcare professionals can provide better decisions for patient diagnosis and treatment. Machine learning algorithms also aid the humans to process huge and complex medical datasets and then analyze them into clinical insights. This paper looks for leveraging deep learning algorithms for detecting a deadly disease, malaria, for mobile healthcare solution of patients building an effective mobile system. The objective of this paper is to show how deep learning architecture such as convolutional neural network (CNN) which can be useful in real-time malaria detection effectively and accurately from input images and to reduce manual labor with a mobile application. To this end, we evaluate the performance of a custom CNN model using a cyclical stochastic gradient descent (SGD) optimizer with an automatic learning rate finder and obtain an accuracy of 97.30% in classifying healthy and infected cell images with a high degree of precision and sensitivity. This outcome of the paper will facilitate microscopy diagnosis of malaria to a mobile application so that reliability of the treatment and lack of medical expertise can be solved.

1. Introduction

Cognitive computing replicates the way humans solve problems while artificial intelligence and machine learning techniques search for creating novel ways for solving problems that humans can potentially do better. A substantial amount of research has been done during the last decades using machine learning algorithms for cost-effective solutions to support healthcare professionals in reducing diseases. Malaria disease originated from Plasmodium parasites through mosquito-borne infection. Malaria is very common over the world mainly in tropical regions. Figure 1 shows how malaria is widely spread across the globe. When infected female Anopheles mosquitoes bite a person, the parasites enter into the blood and begin damaging red blood cells (RBC) that carry oxygen. Flu virus is the malaria's first

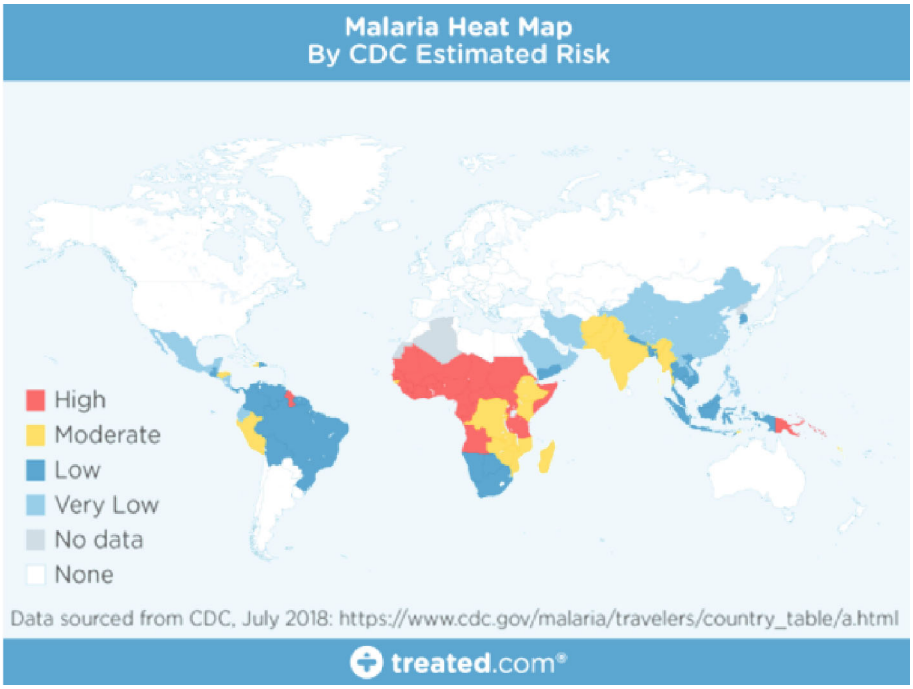


Figure 1: Malaria world map of estimated risk (2018 update) [3].

symptom. The symptom generally starts in few days or weeks. Most importantly, the lethal parasites can stay alive more than a year in a person's body without showing any symptoms. Therefore, a late treatment can cause complications and even death. Hence, many lives can be saved through early malaria detection. Almost 50% of the population in the world is in danger from malaria. There are more than 200 million malaria cases and 400,000 deaths reported every year due to malaria. In practice, to identify malaria, microscopists inspect blood (thick and thin) smears for disease diagnosis and calculate parasitemia. Microscopy examination is used as one of the prime standards for the diagnosis of malaria [1, 2] to identify the existence of parasites in a blood drop from thick blood smears. However, thin blood smears are used for distinguishing the species of parasite and the development of malaria stages. Examination through a microscope is commonly used since it is cheap but time-consuming. The examination accuracy relies on the quality of blood smear and a skilled person who is expert in the classification and examination of uninfected and parasitized blood cells.

Traditional approaches for malaria detection are very time-consuming, may produce inaccurate reports due to human errors, and are laborious for extensive diagnoses. This motivates us to propose an automatic detection of malaria applying deep learning techniques and using a mobile application that leads to early diagnosis which is fast, easy, and effective.

Several ideas exist to detect malaria parasites in microscopic images using convolutional neural networks (CNNs), some pretrained variants of CNN [4–8], and recurrent neural network (RNN) [9]. Moreover, authors in [10, 11] proposed approaches that consider unsupervised machine learning algorithms applying stacked autoencoders for learning the features automatically from the infected and uninfected cell images. Liang et al. [12] proposed a deep learning model for infected malaria cell classification from red blood smears. The model consists of 16-layer convolutional neural network which outperforms transfer learning-based models that use pretrained AlexNet [13].

Jane and Carpenter [14] proposed an object detection-based model using a convolutional neural network, named as Faster R-CNN. The model is first pretrained on ImageNet [15] and then fine-tuned on their dataset. Bibin et al. [16] recommended another model using deep relative attributes (DRA) [17]. Authors use CNN for epilepsy seizure detection [18]. Razzak and Naz [19] have proposed an automated process that considers the tasks of both segmentation and classification of malaria parasites. Their segmentation network consists of a Deep Aware CNN [20], and the classification network employs an extreme learning machine- (ELM-) based approach [21].

Since we are aiming to develop a mobile-based effective solution for malaria detection, we look forward to coming up with a CNN-based deep learning model which is expected to be simpler and computationally efficient in contrast to most of the state-of-the art approaches discussed before that require longer training time. In particular, we make the following contributions: (a) design and evaluation of a base CNN model with standard or no learning schedule and very less trainable parameters to classify parasitized and uninfected cell images, (b) the use of a SGD optimizer with cyclical learning rate schedule along with an automatic learning rate finder in addition to commonly applied regularization techniques in improving the model performance, and (c) deployment of our best performing

model to a mobile application to facilitate simpler and faster malaria detection.

The rest of the paper is organized as follows. Related Work reviews the state-of-the-art techniques used in malaria classification. Materials and Methods provides detailed description of our model, its configuration, dataset used, and performance evaluation metrics. Results and Discussion presents the performance results obtained for our base and improved models and provides state-of-the-art comparison. Finally, Conclusions concludes the paper and outlines some potential future work.

2. Related Work

There has been a significant amount of research during the last decades using computing algorithms for cost-effective solutions to support interoperable healthcare [22] in reducing diseases. For instance, Neto et al. [23] proposed a simulator for simulating events of epidemiology in real time. Kaewkamnerd et al. [24] proposed an image analysis system consisting of five phases for malaria detection and classification. Anggraini et al. [25] developed an application applying image segmentation techniques for separating blood cells' background. Furthermore, Rajaraman et al. [4] implemented feature extractors using pretrained CNN-based deep learning models for uninfected and parasitized blood cell classification to facilitate disease identification. The research used experimental approach to identify the optimal model layers using the underlying data. The CNN model has two fully connected dense layers and three convolutional layers. The performance is measured to extract features using VGG-16, AlexNet, Xception, DenseNet-121, and ResNet-50 from the uninfected and parasitized blood cells. In contrast to [4], only CNN-based malaria classifiers are also proposed by Gopakumar et al. [26] and Liang et al. [12].

MOMALA [27] is a smartphone and microscope-based application developed to detect malaria quickly at a low cost. The MOMALA app can detect the existence of malaria parasites on a regular blood-smeared slide. A phone camera is attached to the microscope's ocular to take the photographs of the blood smear and then analyzes it. At present, the application highly depends on microscopes that are heavy, bulky, and not easily transportable.

The researchers in [28] developed a mobile app that takes photos of blood samples to detect malaria immediately. Using a cell phone app, we can analyze blood samples without involving microscope technicians. The app needs to clamp a smartphone on to a microscope's eyepiece, and the application analyzes the images of the blood sample and creates a red circle on malaria parasites. A lab worker later reviews the case. Extraction of meaningful features is the heart of success for any machine learning method. Most of the computer-used diagnosis tools that use machine learning models for image analysis are based on manual-engineered features for making decision [29–31]. The process also needs computer vision expertise in order to analyze the variability on the images in size, color, background, angle, and position of interests. Deep learning techniques can be applied with considerable success for overcoming the challenges that prevail in a hand-engineered feature extraction process [32]. Models in deep learning apply a series of sequential layers with nonlinear processing hidden units that can find out hierarchical feature relations within the raw image data. The features (low-level) that are abstracted from higher-level features assist in functions of nonlinear decision-making, learning complexity, result in end-to-end extraction of features, and classification [33]. Moreover, deep learning models show better performance compared to kernel-based algorithms such as Support Vector Machines (SVMs), in large volume of data and computational resources, building them to be greatly scalable [34].

A somewhat related pool of work in cognitive computing domain has presented similar contribution. Zhang et al. [35] proposed a protection mechanism for authentication and access control using an interactive robot while controlling private data access stored in cloud. In a subsequent effort [36], they introduced a novel paradigm of cognitive IoT using technologies of cognitive computing. A group of researchers [37] also proposed a module, called Mech-RL, for developing an agent-based literature consultant and a new channel of a meta-path learning method. Furthermore, similar to our battery-operated mobile-based application for malaria detection that can easily be deployed to edge and IoT devices, there is a handful of research [38–41] aiming at developing frameworks on mobile edge to deliver various related services such as secure in-home IoT therapy, content recommendations [42] [43], and position-based services for network amenities [44].

To summarize, the related work mentioned in the literature largely used different pretrained CNN variants such as AlexNet, VGG-16, ResNet-50, Xception, DenseNet-121, and customized CNN models as well for malaria detection in blood smear images and obtained relatively better results than using a custom CNN architecture. However, the downside is that these results are obtained through feature extraction and subsequent training that required long time in some cases [4] a little over 24 hours. In addition, size and complexity of these models make them a bit unrealistic to be used with battery-operated mobile devices. In contrast, we built a simpler and computationally efficient CNN model with considerably less trainable parameters (discussed in Model Configuration section), yet producing comparable or better results keeping in mind our model to be deployed on battery-operated edge and IoT devices such as a smart mobile phone. Moreover, techniques in the literature mostly use de facto SGD optimizer with various learning rate schedules including the adaptive learning rates which suffer from the problem of saddle point or local minima. In contrast, we have used a SGD optimizer with cyclical learning rate schedule along with an automatic optimal learning rate finder which results in faster model convergence with fewer experiments and hyperparameter updates. Finally, most of the state-of-the-art models use image augmentation to increase model generalizability at the expense of longer training time. On the other hand, our model without using data augmentation demonstrates faster convergence and generalizability to unseen data through proper hyperparameter optimization such as learning rate, regularization through batch

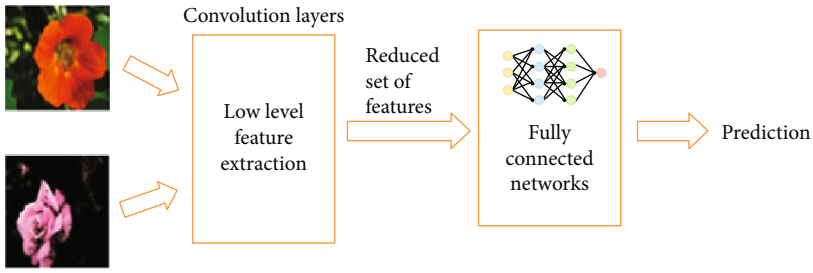
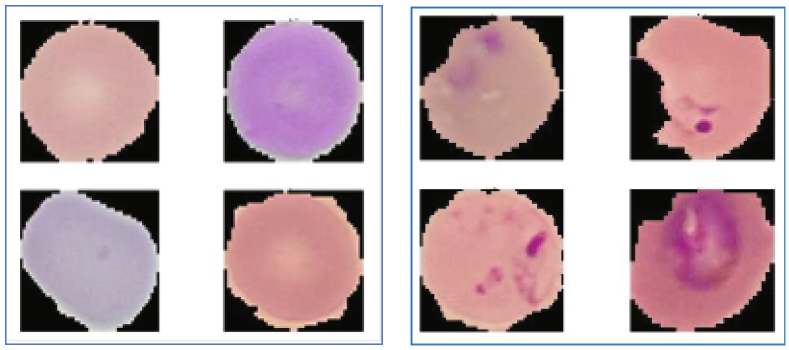


Figure 2: A general CNN model.



(a) (b)

Figure 3: Sample images from NIH dataset: (a) uninfected and (b) parasitized.

normalization, and moderate dropouts in convolutional and dense layers.

Among the studied malaria detection models in the literature, the models proposed in [4, 12, 16, 26] based on custom CNN and its pretrained variants seem to be closest to our model. Hence, we performed a state-of-the-art comparison with these models to demonstrate the feasibility of using our model in a mobile-based system especially in remote disaster survival areas.

3. Materials and Methods

3.1. Deep Learning for Malaria Detection. Deep learning techniques are now widely used for image classification, video recognition, and medical image analysis. A convolutional neural network (CNN), a type of deep neural networks, is mainly considered for research in computer vision field. The deep architecture of CNN is its main power. The convolutional layer in the CNN works as an automatic feature extractor that extracts hidden and important features. Extracted features are passed to a fully connected neural network which performs classification images by maximizing the probability scores. A general CNN model is shown in Figure 2.

3.2. Dataset and Computational Resources. We have used a publicly available malaria dataset from NIH (National Institute of Health) website originally used by a group of researchers, Rajaraman et al. [4], for the detection of malaria parasites in blood smear images. There are 27,558 segmented cell images in the dataset with the same number of normal and parasitized instances. Parasitized cell images contain Plasmodium while normal cells are free of Plasmodium but can contain other staining artifacts and impurities. The data was collected by Chittagong Medical College Hospital in Bangladesh by photographing slides of Giemsa-stained thin blood smear from 200 patients where three-fourth of them were P. falciparum-infected. The manual annotation and deidentification of these collected images were performed by an expert at Mahidol-Oxford Tropical Medicine Research Unit, Thailand, and later approved and archived by Institutional Review Board, National Library of Medicine.

The images in the dataset are not of equal sizes. The minimum and maximum image resolution is 46×46 and 385×395 pixels, respectively, with 3 color channels (RGB). We plan to resize the images to 224×224 which is the standard input image size of the majority of the pretrained CNN models for faster model convergence. Figure 3 shows some sample images from both normal and parasitized categories. The infected cells seem to contain some red globular structures whereas healthy cells do not seem to contain such structures in them. The proposed deep learning model will be used to identify these patterns in cell images to effectively detect malaria parasites in a patient.

Moreover, we performed data scaling which is a crucial preprocessing task for training and evaluating deep learning models. Data from input images without scaling often hampers a steady learning process. Normalization is one of the most common data scaling techniques which rescales the original data points in the images to a range between 0 and 1. The values of data points in the original 8-bit RGB

Table 1: Training and validation data sets.

Set	Count		Percent
	Parasitized	Normal	
Training	11023	11023	80%
Validation	2756	2756	20%

color images range from 0 to 255. Therefore, using Equation (1), we rescale our input image data as follows:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} = \frac{x}{255}. \quad (1)$$

We split the data (as shown in Table 1) into (training and validation) sets randomly with the percentage of 80% and 20%, respectively. There are 22,046 images in the training set and 5512 images in the validation set having an equal number of images from both classes.

The proposed CNN model is trained and evaluated using Google Colab [45] which is a cloud-based Jupyter notebook environment available for free access. Colab provides a preconfigured system for training and evaluating deep learning applications and offers access to high-performance graphical processing units (GPUs) without any cost. Presently, it offers a single 16GB NVIDIA Tesla P100 GPU with CUDA enabled, and all the necessary packages are preinstalled which includes Python 3 with Keras 2.2.5 API and TensorFlow 1.15.0 at the backend. In addition, we have used Android Studio 3.6.1 for developing the android malaria detection app for model deployment.

4. Model Configuration and Evaluation Metrics

4.1. Model Configuration. The proposed CNN model has four convolutional blocks and two fully connected dense layers. Figure 4 shows the proposed CNN model. Each convolutional block consists of convolution, max pooling, batch normalization, and dropout layers. The first convolutional layer uses 32 filters of size 7×7 to learn larger features, and then, the filter size decreases by 2 and filter count is doubled in each subsequent convolutional layer except for the last layer. The default striding of 1 pixel is used in convolution operations in all layers. The model input consists of segmented cell images of $224 \times 224 \times 3$-pixel resolution. In addition, each convolutional layer uses a valid padding to reduce the output feature map dimension in proportionate to the filter size used. We have used nonlinear activation function called ReLU (Rectified Linear Units) in all hidden layers to introduce nonlinearity into the output of each neuron to help the model learn complex mathematical functions to better identify target classes. It removes the vanishing gradient problem and aids in faster model training and convergence [46]. Max pooling layers have a 2×2-pixel pooling window and 2-pixel stride, added after convolutional layers to down sample the feature map by summarizing the most activated existence of a feature. This means that the pooling operation reduces the size of the feature map with a factor of two. To tackle the overfitting problem and to ensure more stability of the network, we have added a batch normalization layer to be applied to pooled output. Normalization is applied on the previous activation layer by subtracting the batch mean and then dividing by standard deviation of the batch [47]. The dropout regularization (with a dropout ratio of 0.15) used in each convolutional block reduces overfitting and improves network generalization error by randomly dropping out nodes during model training [34]. A global average pooling (GAP) layer is considered right after the last block of convolution as a better replacement of flattening to reduce overfitting by minimizing the size of model parameters. The GAP layer reduces spatial dimensions of a 3-dimensional tensor having size $h \times w \times d$ to $1 \times 1 \times d$ tensor by simply taking the average of all hw pixel values of each $h \times w$ feature map to single number [48]. The output from the GAP layer followed by a dropout is passed to the first (fully connected) dense layer having 1000 neurons. The first dense layer output is then fed to a dropout and then passed to the second dense layer with two neurons and a Softmax classifier. Overall, our proposed CNN model has relatively smaller size (409 K) of trainable parameters compared to most of the pretrained transfer learning models used in the literature [4] for malaria detection or solving similar computer vision problems. This simplicity of network structure will be the first step while dealing with overfitting problem.

We consider a stochastic gradient descent (SGD) optimizer with momentum for training and optimizing the model in order to minimize binary cross-entropic loss also known as log-loss. We have optimized our custom model by tuning the learning rate. Learning rate is one of the most dominating hyperparameters in a neural network configuration. We used an automatic optimal learning rate finder in combination with cyclical learning rate (CLR) technique first introduced by Smith [49] which allows the learning rate to oscillate cyclically between a minimum and a maximum learning rate bound. The use of CLR results in faster model convergence with fewer experiments and hyperparameter updates.

4.2. Cyclical Learning Rates. The widely used learning rate schedule technique monotonically decreases learning rate after each epoch to allow the model to descend to a point of low loss. However, with this technique, the model will still be sensitive to initial choice of learning rate and the technique does not guarantee that it will land to a low loss area while decreasing the learning rate. Rather, the model may be confined to either saddle points or local minima. To better address these problems, cyclical learning rates (CLR) enable oscillation of learning rates between upper and lower bounds which in turn provide additional freedom in choosing initial learning rate and get rid of saddle points and local minima. There are three variations of the CLR based on how the oscillation of learning rate takes place: *triangular*, *triangular2*, and *exp_range*. The *triangular* policy works by starting off from a base learning rate and increasing the rate to a maximum value in half cycle and then decreasing back to the initial rate thus completing the full cycle. This whole process is repeated until the model training is finished. The *triangular2* also called triangular schedule with fixed decay is similar to the previous one except that it cuts the

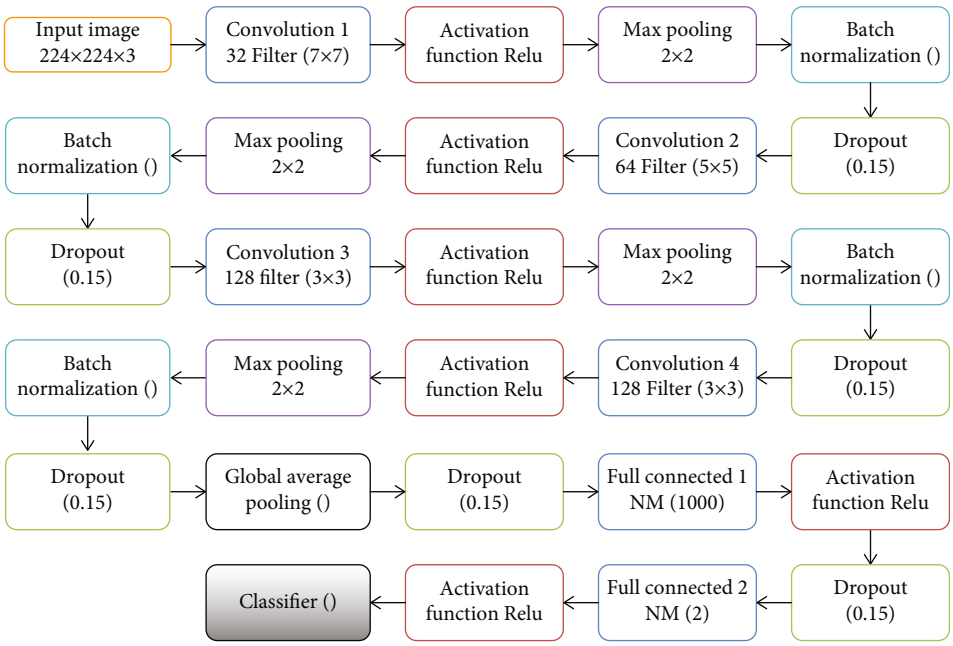


Figure 4: The custom CNN model.

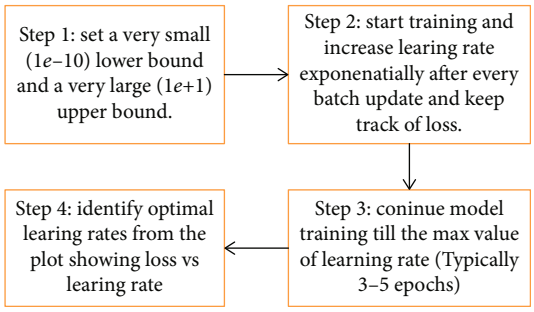


Figure 5: Steps to find optimal learning rates automatically.

upper bound of learning rate to half after every cycle. This lowering of maximum learning rate over time results in increased stability of model training. Finally, *exp_range* policy also called triangular schedule with exponential decay uses an exponential decay as the name suggests to cut down the upper bound which gives more fine-tuned control in decreasing the max learning rate over time. We have used Brad Kenstler's implementation of CLR using Keras for our model training [50].

4.3. Automatic Learning Rate Finder. Since CLR works based on a lower and upper bound of learning rate, Smith [49] also provides an automatic learning rate finder algorithm to find optimal learning rates. Various steps for obtaining the minimum and maximum values of learning rates in Figure 5. For model training, a very small $(1e-10)$ value and a very large $(1e+1)$ value for lower and upper bounds of learning rates are set by the algorithm. An exponential increase of learning rate after every batch update is adopted as training progresses, and at the same time, loss is also recorded. When the learning rate reaches the upper bound after a specific number of training epochs, we plot a curve showing loss and learning rate. At this point, we identify two different values for learning rates. The loss starts decreasing after the first learning rate, and the loss starts to increase from the second value of learning rate. These two values refer to the lower and upper bounds of learning rate which are used in CLR technique. Figure 6 demonstrates how the loss changes with respect to various learning rates using this automatic learning rate finder. It is apparent from the plot that loss does not change until the learning rate hits approximately $1e-6$. This indicates that our model does not start learning owing to a very low initial learning rate. Loss starts to decrease soon after the learning rate reaches approximately $1e-5$ which implies that the learning rate is large enough to enable the model to start learning. From this point, the loss keeps decreasing sharply implying that the model is learning quickly. Soon after the learning rate reaches to approximately $1e-1$, the loss starts to increase again. As such, the loss has exploded almost immediately due to the large increase in learning rate (close to $1e+1$). Hence, we select $1e-5$ and $1e-1$ as our minimum and maximum learning rates, respectively, which will be used in the CLR technique for our model training. Finally, our model configurations including hyperparameters are summarized in Table 2.

4.4. Evaluation Metrics. We have used accuracy, precision, recall, F1-score, specificity, Matthews correlation coefficient (MCC), and Area Under Curve (AUC) to evaluate the performance of our models. Since in our dataset the number of samples from each target class is equal, we consider accuracy as our primary metric. Accuracy refers to the proportion of correct predictions over all predictions made by the model. In addition, we calculated precision, recall or sensitivity, specificity, and F1-score from the confusion matrix which contains False Positives (FP), True Positives (TP), False

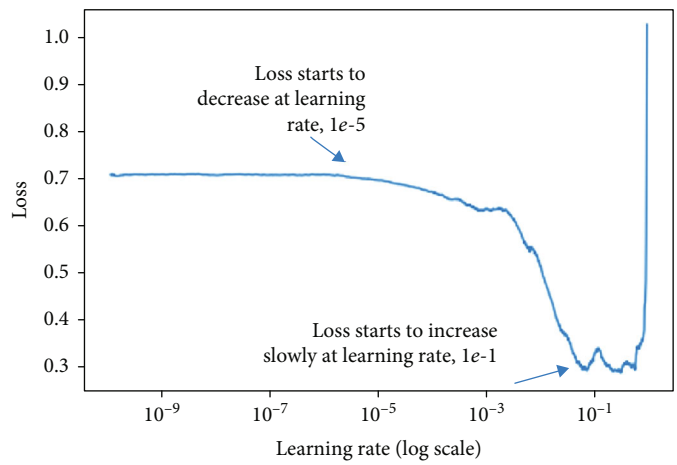


Figure 6: Model loss for various learning rates to identify optimal lower and upper bounds on learning rate.

Table 2: Model configuration summary including hyperparameters.

Parameter	Value/type
Epochs	50
Batch size	32
Optimizer	SGD with momentum 0.9
Learning rates	Min $1e-5$, max $1e-1$
Loss function	Categorical cross entropy
Input shape	224×224
Pooling	Max 2×2 (convolutional layers), GlobalAverage (flatten layer)
Activation	ReLU (convolutional layers), Softmax (final dense layer)
Dropout rate	0.15
Trainable parameters	409,146

Negatives (FN), and True Negatives (TN). Furthermore, precision and recall for each target class are calculated from a classification report. Precision measures the proportion of patients that are identified as infected really carry malaria parasites. Recall or sensitivity measures of the proportion of patients that are infected are diagnosed by the model as having malaria parasites. Specificity is the opposite of recall which measures the proportion of patients that are not infected and diagnosed by the model as not carrying any malaria parasites. F1-score is calculated as a single metric from the harmonic mean of precision and recall. MCC is computed from all four values of confusion matrix and represents the correlation coefficient between the true and predicted classes [51]. The higher the coefficient value, the better is the prediction. Equation (2) is used to calculate MCC for a binary classification problem. When all the predictions of the classifier are correct (i.e., FP = FN = 0), MCC becomes 1 implying the perfect positive correlation. On the contrary, if the predictions are always incorrect (i.e., TP = TN = 0), MCC becomes -1.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \quad (2)$$

We have used binary cross entropy or log-loss, and the target is to minimize it which is equivalent to maximize the classification accuracy. The log-loss function is expressed with the following equation:

$$\mathrm{Loss} = \frac{1}{N}\sum_{i=1}^{N} y_i.\log\left(p(y_i)\right) + (1-y_i).\log\left(1-p(y_i)\right), \quad (3)$$

where y represents the target class (0 for normal cell images and 1 for parasitized cell images) and $p(y)$ is the probability of prediction of the sample being parasitized for all N images. For each parasitized image ($y = 1$), $\log(p(y))$ is added to the loss that is the log probability of its being parasitized. On the contrary, $\log(1 - p(y))$ is added to the loss implying that the log probability of its being normal for each uninfected image ($y = 0$).

5. Results and Discussion

We have adopted the following approach in order to assess the performance of the proposed CNN model for the classification of uninfected and parasitized cell images. We took learning rate as one of the key hyperparameters to tune our custom CNN model for optimum classification performance. The learning rate hyperparameter is used to control the speed of learning of a deep learning model. Using a rightly constructed learning rate, a model can learn to best map input to desired output with the available resources (i.e., the number of nodes in each layer and the total number of layers) with the number of epochs passing in the training data. The SGD algorithm has long been the de facto optimizer to train deep neural networks. Moreover, some of its extensions based on adaptive learning rates have been popular for quite some time now such as Adam, RMSProp, and Adagrad. However, lately, the concept of cyclical learning rates

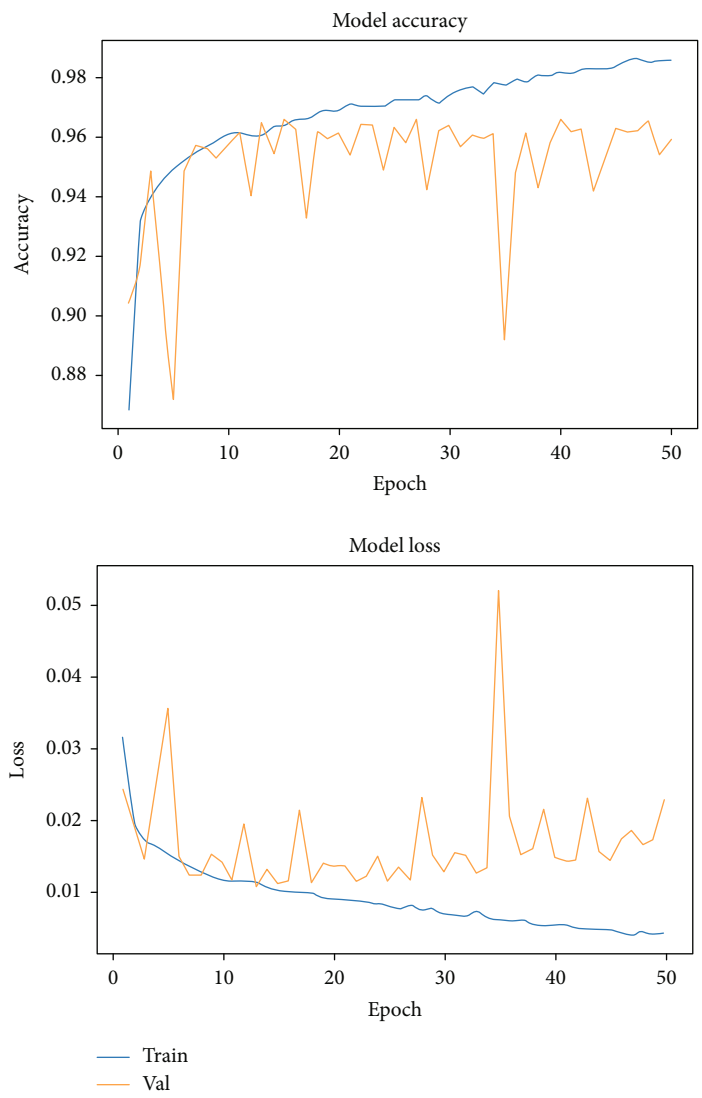


Figure 7: Loss (training and validation) and accuracy of the base model.

(CLR), originally proposed by Smith [49], attracted researchers' attention in improving deep learning model performance. In this paper, we looked at how SGD with this cyclical learning schedule holds up to the other optimizers. To this end, we built a baseline model with standard SGD and then gradually tried to improve the model's performance with CLR technique applied to SGD.

5.1. Base Model. As mentioned in the previous section, our base model represents the custom model as described before with a standard SGD optimizer. We have considered SGD with 0.9 momentum, $1e-1$ as an initial learning rate, and standard decay of *initial_learning_rate/no_of_epochs*. We saved the best model weights (i.e., the lowest validation loss) during training by using Keras's *ModelCheckpoint* library and *callback* function. We trained the model for 50 epochs. Resulted training and validation loss are shown in Figure 7 as well as accuracy over the number of epochs.

We can see that our base model does not converge well and a significant difference between the training and validation results both for loss and accuracy. In addition, a lot of fluctuation is observed in the values of loss and accuracy as the training progresses towards the end. This indicates that our base model is not trained well and might be overfitting to training data. Consequently, the model might not generalize well on unseen test data. This could be potentially attributed to the choice of learning schedule in the base model even though we have used dropout and batch normalization techniques to avoid overfitting. We aim to overcome these drawbacks by using CLR schedule with the SGD optimizer.

Table 3: Performance metrics for the base model.

Acc	AUC	Precision	Recall	F1-score	MCC
0.9646	0.9552	0.97	0.96	0.96	0.9135

Performance metrics of our base model is shown in Table 3. We have received a base accuracy of 96.46% with high precision and recall towards classifying the infected

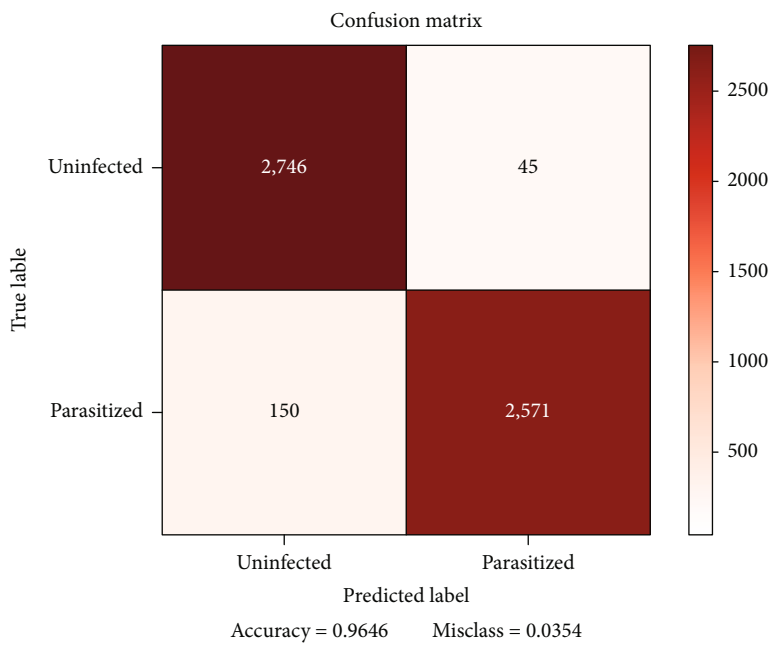


Figure 8: Confusion matrix for the *base model*.

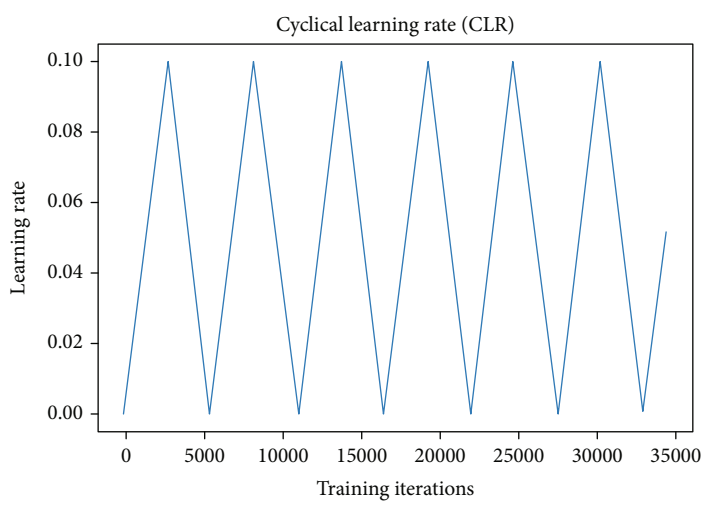


Figure 9: Cyclical learning rate changes using "triangular" policy. Lower and upper bounds on learning rates were calculated using an automatic learning rate finder.

and normal cells which is reasonable. By investigating the confusion matrix as shown in Figure 8, we can see that the count for False Negatives (FN) is 150 which is pretty high for a disease identification problem. FN indicates that the model declares a malaria patient to be healthy whereas the patient is parasitized. This will severely hamper the patient treatment and may result in death. Our goal is to reduce this number with the proposed improved model. A reduced number of FN will ensure that our model is effective in identifying parasitized cell images.

5.2. Improved Model with CLR. In our improved model, we used the same base model architecture with the exception that we used cyclical learning rates schedule instead of a standard one. As mentioned before, there are three variations of CLR implementation based on the policy of changing the upper bound learning rate, namely, *triangular*, *triangular2*, and *exp_range*. We have experimented with the first two variants to observe the model performance. Figure 9 shows the learning rate plot and how the learning rates oscillate between the lower and upper bounds.

More specifically, the initial learning rate of $1e-5$ increases to the maximum value of $1e-1$ in a half cycle and then decreases back to $1e-5$ in the other half cycle thus completing the full cycle. By using this triangular policy, we have obtained improved model accuracy of 97.12% as shown in Table 4 (compared to 95.25% in base model) with higher precision and recall towards classifying the infected and

Table 4: Performance metrics for the improved model with CLR schedule.

Model	Base	CLR-triangular	CLR-triangular2
Accuracy	0.9646	0.9712	**0.9730**
AUC	0.9552	0.9656	**0.9704**
Precision	0.97	0.97	**0.97**
Recall	0.96	0.97	**0.97**
F1-score	0.96	0.97	**0.97**
MCC	0.9135	0.9400	**0.9417**

normal cells. Thus, by combining cyclical learning rates with the automatic learning rate finder (discussed earlier), we are successful in obtaining a highly accuracy model.

By looking at the training history (plotted in Figure 10), we found that the gap between training and validation loss as well as accuracy reduces significantly indicating a faster and better model convergence. In addition, we observed a "wave" characteristic of our training and validation accuracy/loss curve signifying the fact that the learning rate oscillates between lower and upper bounds.

We train and evaluate our model again with "triangular2" CLR policy and found out further improvement in model accuracy which is 97.30%. Figure 11 visualizes how learning rate is adapted in a cyclic manner and, after each fully cycle, the upper bound learning rate is reduced to half and this continues till the end of model training. Compared to the first triangular CLR policy, the training and validation curves with loss and accuracy (as plotted in Figure 12) show less fluctuation and more stability. In principle, a stabilized training is less prone to the risk of overfitting.

Table 5 shows the confusion matrix for our improved model with *triangular2* CLR schedule. As mentioned earlier, our target is to reduce the FN count to make our model robust. We can see the FN count decreased to 112 compared to the base model (with FN count of 150) which makes our improved model effective in identifying parasitized cell images. This reduced count of FN is very critical because we do not expect that our model will misidentify someone as healthy while in reality the patient is carrying the malaria parasite. This will severely hamper the patient's line of treatment and even endanger life of the patient. At the same time, FP count also decreased (to 37) compared to the base model (45). A lower value FP is also expected from our model since this will prevent the patient from further undergoing unnecessary laboratory tests and treatment and will reduce financial burden on the health provider.

5.3. Mobile-Based Model Deployment. We have deployed our best improved model to a mobile application to facilitate a simple and fast detection of malaria parasite in blood cell images. We have used Google's TensorFlow Lite [52] which brings deep learning capability directly into mobile devices by running deep learning models locally. TensorFlow Lite framework supports hardware acceleration and brings low-latency inference performance to mobile devices by significantly improving model loading times. Figure 13 shows different steps of our model deployment process. Our best trained Tensorflow model is converted to a TensorFlow Lite (*.tflite*) model using the TFLite Converter. We have used the TFLite Converter from a Python API which simplifies the model conversion as part of a model deployment pipeline. Our converted *.tflite* mode size is about 22 MB. Once the converted (*.tflite*) model is deployed on the android mobile device, cell images are loaded from a cloud or device's local storage for potential malaria detection. The user opens a cell image, and the deployed model provides the prediction label. Snapshots of a few sample image predictions are displayed in Figure 14.

6. Discussion

From the performance results obtained in the previous section, we observed that the proposed custom model with *CLR-triangular2* configuration produces an optimal solution with faster convergence. This was achieved by selecting a superior combination of convolutional and dense layers in the custom CNN architecture with proper hyperparameter optimization such as learning rate, regularization through batch normalization, and moderate dropouts in convolutional and dense layers. The use of cyclical learning rate schedule with an automatic learning rate finder lowered the effect of model being overfit to training data and faster convergence to a better solution.

Our base model with no or standard learning rate schedule did not converge well and showed high variance in the values of loss and accuracy during the model training indicating a tendency to overfit to training data. We have addressed this problem by using cyclical learning rate schedule in our SGD optimizer along with implicit regularization techniques using batch normalization and dropouts. The use of two different variations of cyclical learning rate implementation, namely, triangular with no decay (triangular) and triangular with fixed decay (trianglular2), progressively improves the performance of the base model with respect to model accuracy, AUC, sensitivity (recall), and MCC. Our best improved model yields a performance accuracy of 97.30% compared to the base model's accuracy of 95.57%. A noticeable increase in the value of MCC (94.17% from the base model's MCC of 91.35%) indicates that the predicted label and the true label are strongly correlated and our improved model is competent in classifying parasitized and uninfected cell images. An increased value of AUC (97.04%) represents a high degree of separability of our improved model meaning that it can better distinguish between cell images with malaria disease and no disease. In addition, a high recall value of 97% indicates model sensitivity in predicting infected cells with malaria. Furthermore, the number of false negative (FN) cases significantly (almost half) decreased in our best improved model compared to the base model which indicates the success of our model in reducing the risk of identifying a malaria patient as healthy which is detrimental to a patient's line of treatment. It is worth mentioning here that our model did not consider using data augmentation to artificially increase the size of the training dataset largely owing to the fact that our dataset contains a decent number (27,558) of segmented cell images with the same number of normal

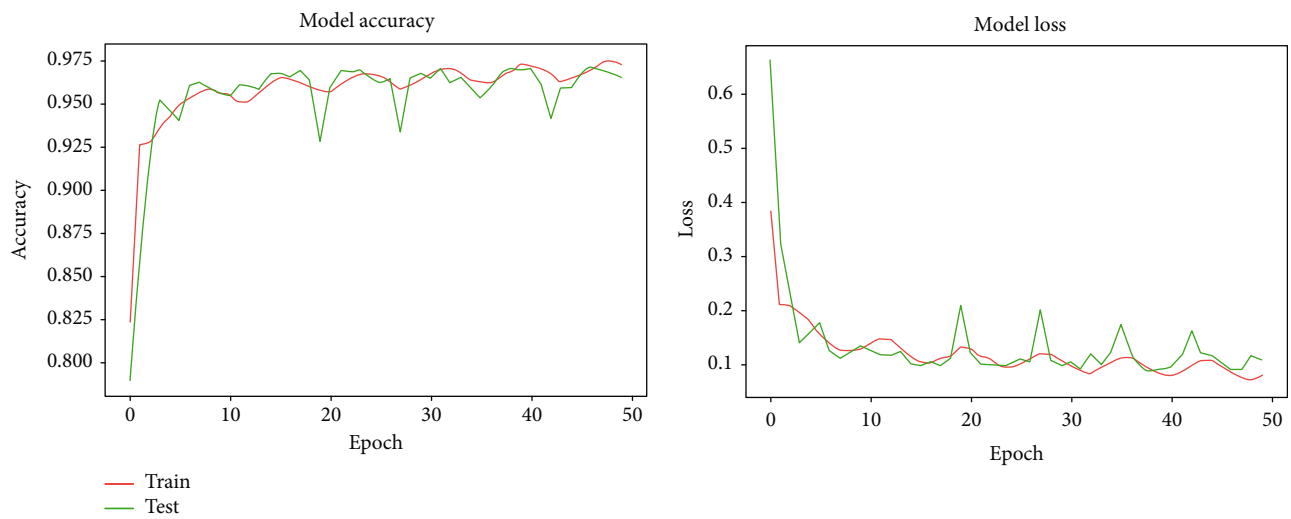


Figure 10: Training and validation loss and accuracy with *triangular* CLR policy.

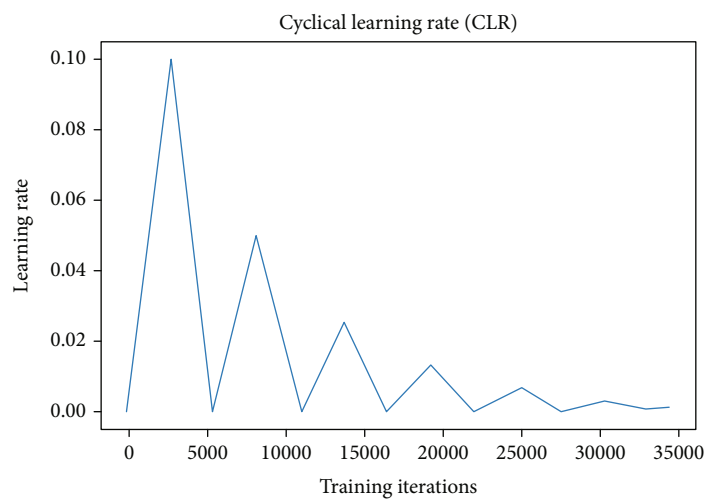


Figure 11: Cyclical learning rate changes using "triangular2" policy.

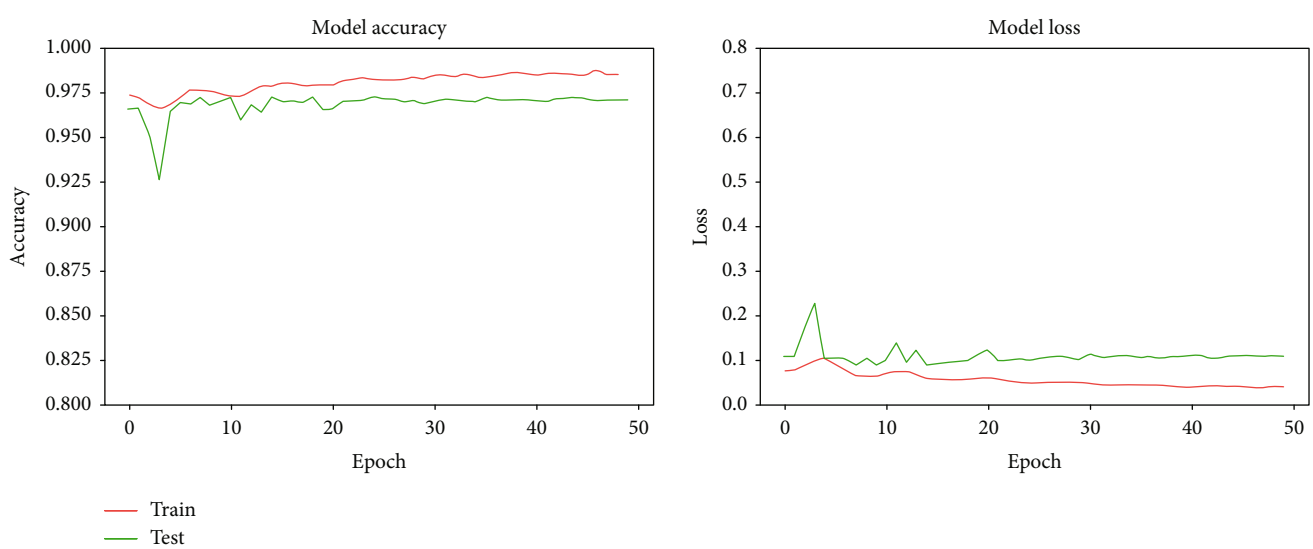


Figure 12: Training and validation loss and accuracy with *triangular2* CLR policy.

Table 5: Confusion matrix for the improved model with *triangular2* CLR schedule.

Model	TP	TN	FN	FP
Base	2746	2571	150	45
CLR-triangular	2752	2601	120	39
CLR-triangular2	2754	2609	112	37

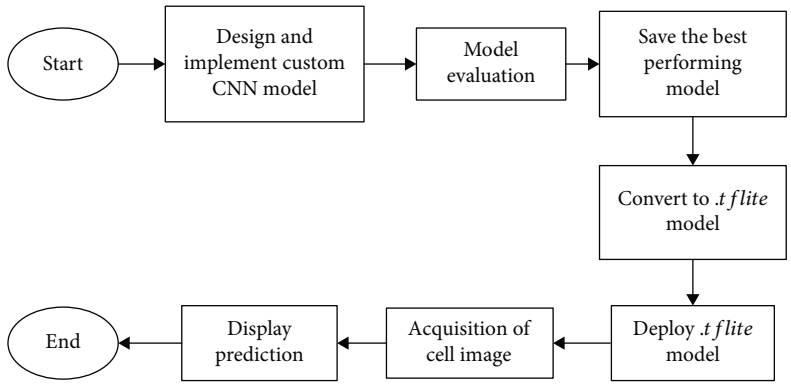


Figure 13: Process flow diagram showing different steps of model deployment in a mobile device.

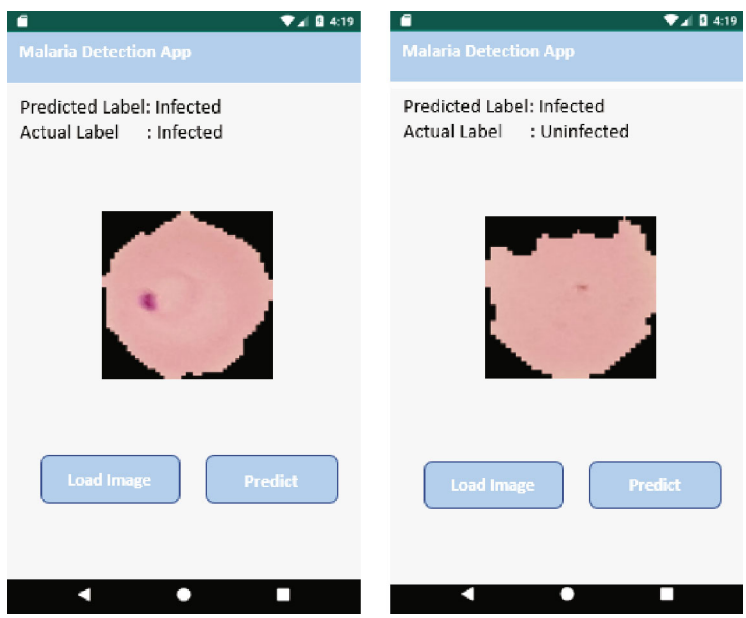


Figure 14: Snapshots of the mobile app displaying predictions on actual cell images.

and parasitized instances well enough to train a deep learning model without significantly running into overfitting problem. Hence, we trained our model without artificially augmenting our dataset and yet obtained better or comparable performance to the techniques in the literature in identifying a malaria patient.

Table 6 provides a comparison of performance metrics between our best improved model and the results of state-of-the-art approaches. We noticed that the proposed improved model is better than the customized model and other CNN models (pretrained) such as VGG-16 and ResNet-50 presented in [4] with respect to accuracy, precision, sensitivity, and MCC towards classifying healthy and infected cells with malaria. On the contrary, our proposed custom model achieved a relatively lower value for AUC as compared to ResNet-50 and VGG-16 but demonstrated similar AUC performance as the customized model proposed in [4]. Our model took about 97 min to train as compared to the training time (24 hours) of all the models proposed in [4]. We believe that this performance improvement is worth given the fact that our model is smaller in size having a relatively less number of trainable parameters and demonstrated very less training time using the SGD optimizer with a cyclical learning rate schedule.

Table 6: Comparison of performance of proposed and the state-of-the-art approaches.

Model	Accuracy	AUC	Precision	Recall (sensitivity)	F1-score	MCC
Proposed model (CLR-triangular2)	0.9730	0.9704	0.97	0.97	0.970	0.9417
Rajaraman et al. Customized model [4]	0.9400	0.9790	0.951	0.931	0.941	0.880
Rajaraman et al. ResNet-50 [4]	0.9570	0.9900	0.969	0.945	0.957	0.912
Rajaraman et al. VGG-16 [4]	0.9450	0.9810	0.951	0.939	0.945	0.887
Gopakumar et al. [26]	0.9770	—	0.985	0.971	0.977	0.731
Bibin et al. [16]	0.963	—	0.959	0.976	0.967	—
Liang et al. [12]	0.973	—	0.977	0.969	0.972	—

In contrast to the pretrained models presented in [4], the CNN-based classifier proposed by Gopakumar et al. [26] demonstrated slightly better results in classifying parasitized and uninfected cells in terms of accuracy (97.70%), precision (98.5%), and recall (97.1%) but showed a very low MCC (73.1%) value which is considered a very informative consolidated score for evaluating a binary classifier's performance representing the correlation between the predicted and true classes [51]. Liang et al. [12] have also proposed a technique for image analysis using a CNN for malaria detection. They have achieved similar accuracy (97.3%) as our improved model with a slight increase in precision (97.7%) and slightly degraded sensitivity (96.9%) as compared to our improved model. Finally, malaria parasite detection using a deep belief network done by Bibin et al. [16] did not demonstrate promising results as compared to other studies in the literature including our improved model. Based on the preceding discussion, our model is greatly specific with a large MCC value and performs pretty better than the majority of the pretrained and custom CNN models under study.

7. Conclusions and Future Work

The paper first evaluated a custom CNN-based end-to-end deep learning model to improve malaria detection on thin-blood smear images. We showed that the use of cyclical learning rate schedule with an automatic learning rate finder in addition to the use of a commonly applied regularization technique such as batch normalization and dropouts produces promising results in malaria classification. Our best model achieves an accuracy of 97.30% in classifying parasitized and uninfected cell images with a high degree of precision and sensitivity. The model also yields a high value of MCC (94.17%) compared to all other existing models under study indicating a strong correlation between predicted and true labels. We also observed that the proposed improved model showed better performance compared to the customized and other CNN models (pretrained such as VGG-16 and ResNet-50) [4] with respect to accuracy, precision, sensitivity, and MCC towards classifying healthy and infected cells with malaria. We deployed our best performing model into an android-based mobile application to facilitate simpler and faster malaria detection. Thus, we believe that the results obtained from this work will benefit towards developing valuable mobile-based solutions so that reliability of the treatment and lack of medical expertise can be solved. As an immediate extension of this work, we will consider using image augmentation on the training data with the hope to further alleviate overfitting problem and different adaptive variants of the SGD optimizer to observe their impact on the performance results. In the future, we also plan to achieve better prediction by using ensemble methods through model stacking.

Data Availability

Data set is collected from Kaggle (https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria) [53].

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This study was funded by the Deanship of Scientific Research, Taif University, KSA (Research Project number: 1-440-6146).

References

[1] K. S. Makhija, S. Maloney, and R. Norton, "The utility of serial blood film testing for the diagnosis of malaria," *Pathology*, vol. 47, no. 1, pp. 68–70, 2015.

[2] WHO, *Malaria Micropscopy Quality Assurance Manual*, World Health Organization, 2016.

[3] "Our Malaria World Map of Estimated Risk (2018 update)," https://www.treated.com/malaria/world-map-risk.

[4] S. Rajaraman, S. K. Antani, M. Poostchi et al., "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," *PeerJ*, vol. 6, article e4568, 2018.

[5] C. Mehanian, M. Jaiswal, C. Delahunt, C. Thompson, M. Horning, and L. Hu, "Computer-automated malaria diagnosis and quantization using convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 116–125, Venice, Italy, 2017.

[6] E. Var and F. B. Tek, "Malaria parasite detection with deep transfer learning," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pp. 298–302, Sarajevo, Bosnia-Herzegovina, September 2018.

[7] A. Vijayalakshmi and B. Rajesh Kanna, "Deep learning approach to detect malaria from microscopic images," *Multimedia Tools and Applications*, vol. 79, 2019.

[8] Y. Souri, E. Noury, and E. Adeli, "Deep relative attributes," in *Computer Vision – ACCV 2016*, pp. 118–123, Springer, 2016.

[9] M. I. Razzak, "Malarial parasite classification using recurrent neural network," *Journal of Image Processing (IJIP)*, vol. 9, no. 2, 2015.

[10] H. Shen, W. D. Pan, Y. Dong, and M. Alim, "Lossless compression of curated erythrocyte images using deep autoencoders for malaria infection diagnosis," in *2016 Picture Coding Symposium (PCS)*, pp. 1–5, Nuremberg, Germany, December 2016.

[11] I. Mohanty, P. A. Pattanaik, and T. Swarnkar, "Automatic detection of malaria parasites using unsupervised techniques," in *International Conference on ISMAC in Computational Vision and Bio-Engineering*, pp. 41–49, Springer, Cham, 2018.

[12] Z. Liang, A. Powell, I. Ersoy et al., "CNN-based image analysis for malaria diagnosis," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 493–496, Shenzhen, China, December 2016.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097–1105, 2012.

[14] H. Jane and A. Carpenter, "Applying faster R-CNN for object detection on malaria images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 56–61, Honolulu, HI, USA, 2017.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, June 2009.

[16] D. Bibin, M. S. Nair, and P. Punitha, "Malaria parasite detection from peripheral blood smear images using deep belief networks," *IEEE Access*, vol. 5, pp. 9099–9108, 2017.

[17] X. Yang, T. Zhang, C. Xu, S. Yan, M. S. Hossain, and A. Ghoneim, "Deep relative attributes," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1832–1842, 2016.

[18] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying deep learning for epilepsy seizure detection and brain mapping visualization," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1s, pp. 1–17, 2019.

[19] M. I. Razzak and S. Naz, "Microscopic blood smear segmentation and classification using deep contour aware CNN and extreme machine learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 801–807, Honolulu, HI, USA, July 2017.

[20] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: context-aware deep network models for weakly supervised localization," in *Computer Vision – ECCV 2016*, pp. 350–365, Springer, 2016.

[21] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[22] M. Masud, M. S. Hossain, and A. Alamri, "Data interoperability and multimedia content management in e-health systems," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1015–1023, 2012.

[23] O. B. Leal Neto, C. M. Albuquerque, J. O. Albuquerque, and C. S. Barbosa, "The schisto track: a system for gathering and monitoring epidemiological surveys by connecting geographical information systems in real time," *JMIR Mhealth Uhealth*, vol. 2, no. 1, article e10, 2014.

[24] S. Kaewkamnerd, C. Uthaipibull, A. Intarapanich, M. Pannarut, S. Chaotheing, and S. Tongsima, "An automatic device for detection and classification of malaria parasite species in thick blood film," *BMC Bioinformatics*, vol. 13, Supplement 17, p. S18, 2012.

[25] D. Anggraini, A. S. Nugroho, C. Pratama, I. E. Rozi, A. A. Iskandar, and R. N. Hartono, "Automated status identification of microscopic images obtained from malaria thin blood smears," in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, Bandung, Indonesia, July 2011.

[26] G. P. Gopakumar, M. Swetha, G. S. Siva, and G. R. K. Sai Subrahmanyam, "Convolutional neural network-based malaria diagnosis from focus stack of blood smear images acquired using custom-built slide scanner," *Journal of Biophotonics*, vol. 11, no. 3, 2018.

[27] "MOMALA," https://momala.org/malaria-diagnosis/.

[28] "This New App Helps Doctors Diagnose Malaria in Just 2 Minutes," https://www.globalcitizen.org/en/content/app-diagnose-malaria-uganda.

[29] N. E. Ross, C. J. Pritchard, D. M. Rubin, and A. G. Dusé, "Automated image processing method for the diagnosis and classification of malaria on thin blood smears," *Medical & Biological Engineering & Computing*, vol. 44, no. 5, pp. 427–436, 2006.

[30] D. K. Das, M. Ghosh, M. Pal, A. K. Maiti, and C. Chakraborty, "Machine learning approach for automated screening of malaria parasite using light microscopic images," *Micron*, vol. 45, pp. 97–106, 2013.

[31] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. R. Thoma, "Image analysis and machine learning for detecting malaria," *Translational Research*, vol. 194, pp. 36–55, 2018.

[32] M. S. Hossain, M. Al-Hammadi, and G. Muhammad, "Automatic fruit classification using deep learning for industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 1027–1034, 2019.

[33] M. Usama, B. Ahmad, J. Wan, M. S. Hossain, M. F. Alhamid, and M. A. Hossain, "Deep feature learning for disease risk assessment based on convolutional neural network with intra-layer recurrent connection by using hospital big data," *IEEE Access*, vol. 6, pp. 67927–67939, 2018.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[35] Y. Zhang, Y. Qian, D. Wu, M. Shamim Hossain, A. Ghoneim, and M. Chen, "Emotion-aware multimedia Systems security," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 617–624, 2019.

[36] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive internet of things: improving sensitivity and interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58–64, 2019.

[37] X. Ma, R. Wang, Y. Zhang, C. Jiang, and H. Abbas, "A name disambiguation module for intelligent robotic consultant in

industrial internet of things," *Mechanical Systems and Signal Processing*, vol. 136, article 106413, 2020.

[38] Y. Zhang, M. S. Hossain, A. Ghoneim, and M. Guizani, "COCME: content-oriented caching on the mobile edge for wireless communications," *IEEE Wireless Communication*, vol. 26, no. 3, pp. 26–31, 2019.

[39] J. Wang, Y. Miao, P. Zhou, M. S. Hossain, and S. M. M. Rahman, "A software defined network routing in wireless multihop network," *Journal of Network and Computer Applications*, vol. 85, pp. 76–83, 2017.

[40] A. K. Sangaiah, D. V. Medhane, T. Han, M. S. Hossain, and G. Muhammad, "Enforcing position-based confidentiality with machine learning paradigm through mobile edge computing in real-time industrial informatics," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4189–4196, 2019.

[41] M. A. Rahman, M. M. Rashid, M. Shamim Hossain, E. Hassanain, M. F. Alhamid, and M. Guizani, "Blockchain and IoT-Based Cognitive Edge Framework for Sharing Economy Services in a Smart City," *IEEE Access*, vol. 7, pp. 18611–18621, 2019.

[42] M. F. Alhamid, M. Rawashdeh, H. Al Osman, M. S. Hossain, and A. El Saddik, "Towards context-sensitive collaborative media recommender system," *Multimedia Tools and Applications*, vol. 74, no. 24, pp. 11399–11428, 2015.

[43] Y. Zhang, Y. Li, R. Wang, M. S. Hossain, and H. Lu, "Multi-Aspect Aware session-based recommendation for intelligent transportation services," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2020.

[44] Y. Zhang, R. Wang, M. S. Hossain, M. F. Alhamid, and M. Guizani, "Heterogeneous information network-based content caching in the internet of vehicles," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 10216–10226, 2019.

[45] "Google Colab," https://colab.research.google.com/.

[46] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," in *Proc of 33rd international conference on machine learning (ICML2016)*, vol. 48, pp. 2217–2225, New York, USA, 2016.

[47] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. of the 32nd International Conference on Machine Learning*, vol. 37, pp. 448–456, Euralille Lille, France, 2015.

[48] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. of International Conference on Learning Representations (ICLR)*, Scottsdale, AZ, USA, 2013https://arxiv.org/abs/1312.4400.

[49] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, Santa Rosa, CA, USA, March 2017.

[50] B. Kenstler, "Cyclical Learning Rates Implementation," https://github.com/bckenstler/CLR.

[51] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[52] "TensorFlow Lite- Deploy machine learning models on mobile and IoT devices," https://www.tensorflow.org/lite.

[53] "Malaria Cell Images Dataset," https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8888296, 11 pages
https://doi.org/10.1155/2020/8888296

Research Article

Research on Privacy Security Risk Assessment Method of Mobile Commerce Based on Information Entropy and Markov

Tao Zhang,[1] Kun Zhao,[1] Ming Yang,[1] Tilei Gao,[1] and Wanyu Xie[2]

[1]*School of Information, Yunnan University of Finance and Economics, Kunming 650221, China*
[2]*Personnel Department, Kunming Metallurgy College, Kunming 650033, China*

Correspondence should be addressed to Ming Yang; yangming@ynufe.edu.cn

Received 31 May 2020; Revised 11 June 2020; Accepted 17 June 2020; Published 7 July 2020

Academic Editor: Yin Zhang

To obtain precise personalized services in mobile commerce, the users have to disclose their personal information to the operator, which constitutes a potential threat to their privacy security. In this paper, a mobile commerce privacy security risk assessment model is established based on information entropy and Markov chain, and effective security risk measurement, and assessment method is put forward. Our method can provide accurate and quantitative results in assessing privacy disclosure risk to guide the users' selection of safe mobile commerce applications and protect their privacy security.

1. Introduction

In the mobile internet age, mobile commerce (m-commerce for short) has gained a high market share by virtue of its portable characteristics, and various precise services like web access, e-shopping, tourism consumption, and near-field payment are rendered to the public. With the popularization of m-commerce, the users can access more and more precise services, but meanwhile, their privacy and security are facing serious threats [1]. To obtain and enjoy more precise personalized services, the users have to disclose more personal information to the service operator, and the operator requires more details of such information to maintain the operation of the commercial platform and render the so called diverse personalized services. Then, the private information of users may be disclosed, abused, stolen, or exposed to other risks when being acquired, used, transmitted, and stored by the operator, and multiple data, including social security number, credit card number, protected health information, and user name, may be disclosed unintentionally. Meanwhile, the private information can also be stolen via internal theft, external hacking, employee negligence, or in other ways. As learned by the Identity Theft Resource Center and the US Department of Health and Human Services, the top 10 data breaches of 2019, where more than 137 million records were leaked, were all related to the government, medical institutions, and corporate websites or apps [2]. The academia and industry are paying more attention to the security risk of users' private information in mobile commerce.

At present, most researchers focus on the risk assessment of private information in information system, cloud computing, and big data, and the risk assessment of user private information disclosure in m-commerce is rarely studied. Given the vital importance of risk assessment for information security to the ecosystem and sustainable development of m-commerce platform [3], the risks of users' private information in m-commerce are explored in this paper from the perspective of private information disclosure. Compared with the traditional information system security risk factors, the risk hierarchy structure of users' privacy information in m-commerce is more complex. These risks include traditional information system security risk, user behavior risk, third-party application risk, and special risks of m-commerce services, like the risk in location-based services in mobile networks [4]. Therefore, in this paper, various risk factors are comprehensively analyzed by reference to some literature, and a risk indicator system for user private information disclosure in m-commerce is built based on the security model of information system [5]. Moreover, the privacy security of users is still assessed, and effective risk assessment

model is built based on the theories of information entropy and Markov chain, to provide accurate risk assessment results to the users and protect their privacy security in m-commerce.

This paper can divided into the following parts: in Section 1, the background, content, and significance of the research are presented; in Section 2, we summarize and discuss the privacy security risk index, measurement and assessment methods in m-commerce are summarized and expounded, and the existing problems in the current researches on privacy security of m-commerce are revealed; in Section 3, we apply information entropy and Markov chain in the research of privacy security risk of m-commerce users, the user privacy security is described based on the information entropy, and the random state of privacy security risk of m-commerce is restored in accordance with Markov chain; in Section 4, a risk assessment model for m-commerce user privacy disclosure is established based on information entropy and Markov chain, effective assessment method is put forward, and the whole assessment process is specified; in Section 5, a detailed case study is carried out by substituting the proposed model into a specific m-commerce application, and the quantitative assessment results for three applications are presented and compared with each other. And finally, in Section 6, the research of this paper is summarized, and the future research direction is pinpointed.

2. Related Work

Recent researches on privacy security risk of m-commerce can be generally classified into two aspects, identification of risk factor and method development for risk assessment.

2.1. Research on Risk Factors of User Privacy Disclosure. Risk assessment depends on the identification of risk factors. In order to properly define the privacy risks of m-commerce users, we conclude the risk factors that have been widely studied by researchers in Table 1.

Table 1: Classes and factors of privacy risks of m-commerce users.

Risk classes	Risk factors
Technology risk	Data encryption; intrusion detection; authorization and authentication; access control; anonymisation; trajectory information hiding
Platform environmental risk	Data contribution agreement; secure routing protocol; legal or institutional requirements; diversity of privacy laws; mobile advertising attack; location service
Operator management risk	Privacy management mechanism; regulatory and disciplinary systems; insider threat; third party information collection
User vulnerability risk	Privacy awareness; privacy invasion experience; privacy association setting; simple password setting
Mobile terminal device risk	Sensitive data protection; taint tracking; privilege manage; detection of malicious events

2.1.1. Technology Risk. Shirazi and Iqbal [6] studied the community clouds in m-commerce and pointed out that the privacy security of users in m-commerce mainly relies on data encryption, intrusion detection, identity management, security awareness, privacy protocol, privacy principle, privacy practice, and effective database utilization. Erfan et al. [7] suggested that anonymous technology could help reduce the personal privacy risk of m-commerce users. Zhang et al. [8] proposed a security policy based on identity authentication and access control to protect private information stored in the edge cloud. Yosef and Mahmoud [9] analyzed the security issues at various levels of the cyber physical system (CPS) architecture and pointed out that to improve its safety, attention should be paid to the influence of relevant technologies, such as authentication, access control, data encryption, environment monitoring, security routing protocol, network access control, attack detection mechanism, and user authentication and authorization.

2.1.2. Platform Environmental Risk. According to literature [10, 11], location information was extremely sensitive in m-commerce, and the exposure of location information might cause the risk of information abuse in m-commerce. In reference [12], it was found that advertisements in m-commerce were intrusive to the users' privacy, for the users' location, and other information may be mandatorily acquired. Reference [13] reveals that users are required to accept some privacy clauses before using some m-commerce applications and have no autonomy over whether to share their own information in utilization.

2.1.3. User Vulnerability Risk. Ampong et al. [14] noted that privacy awareness, privacy concerns, and privacy intrusion experiences were important factors that affected the disclosure of user privacy. Reference [15] conducted a qualitative analysis of the privacy risk factors of social networks in the big data environment and suggested that privacy association setting, spatial location sharing, information behavior negligence, and simple password setting constituted the major user behavior risk factors.

2.1.4. Operator Management Risk. Tian et al. [16] believed that the privacy risks in the management of mobile apps included rigid legal or institutional requirements, imperfect standards for disclosure of privacy information, lack of regulatory and disciplinary systems, and malicious disclosure by internal personnel. In line with the Risk Evaluation Specification for Information Security (GB-T20984-2007) and the behavior characteristics of m-commerce users, Xiang et al. [17] incorporated into their risk evaluation index system such related factors as privacy management mechanism, platform privacy protection input, information sharing risk, third party information collection, and privacy legal differences.

2.1.5. Mobile Terminal Device Risk. In addition to the risk factors mentioned above, potential privacy risk may arise from the mobile terminal as well. Therefore, corresponding measures, including sensitive data protection [18, 19], smear

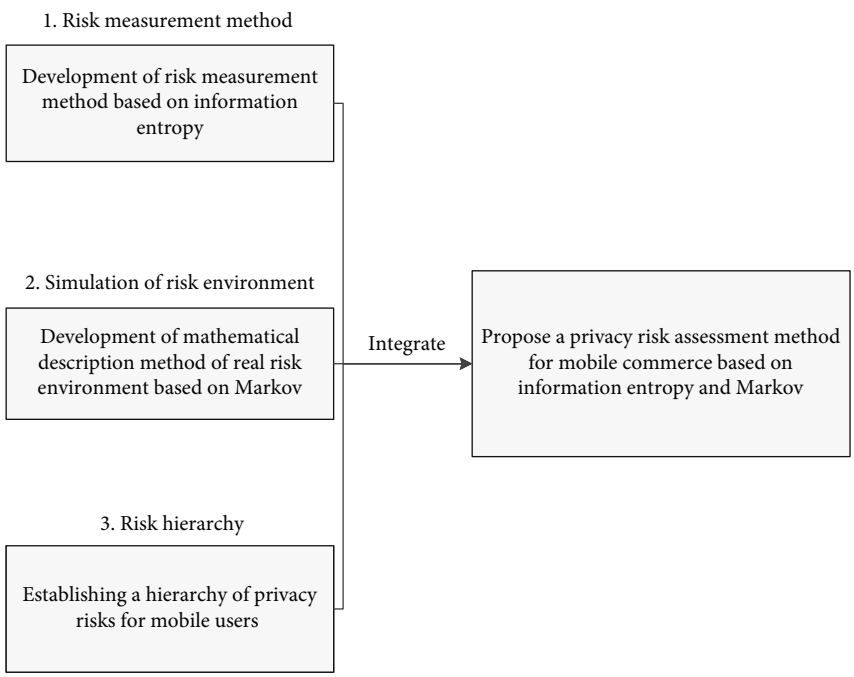


Figure 1: Framework of our research in developing risk assessment method.

tracking [20], authority management [21], and malicious event monitoring [22], need to be taken to ensure the security at the mobile terminal.

2.2. Research on Assessment Method for Privacy Security Risk. At present, fruitful achievements have been made in the research of risk assessment, but only a few researches focus on privacy risk assessment, and researches on privacy assessment for m-commerce applications are rare. In references [23–26], the risks were evaluated based on the concept of information entropy; a feasible program was proposed for the assessment of security risk in cloud computing, but the privacy security was not analyzed. In reference [27], a privacy-considered information security assessment model was built with the risk recommendation system based on the identifiability, context of use, quantity, sensitivity, and freshness of the personal identity information data. The likelihood of risk evaluation was calculated taking into account the impact assessment of existing control measures and risks, and privacy security was evaluated from the perspective of the frequency of risk occurrence. Oetzel and Spiekermann [28] proposed a system approach for privacy impact evaluation, and divided the entire privacy impact assessment (PIA) process into seven steps, namely, characterization of the system, definition of privacy objectives, evaluation of protection requirements, identification of threats, identification and recommendations of controls, evaluation of residual risks, and PIA documentation. Taking into account the new challenges of user privacy management, Lo et al. [29] worked out LRPdroid, a user privacy analysis framework for the Android platform, to detect the information leak and evaluate user privacy leak and privacy risks for applications installed on android-based mobile devices. These methods have significant reference value for the risk assessment of this paper. However, only a certain class of privacy security risk was evaluated with above methods, taking into account neither the interaction between various risks nor the risk characteristics of m-commerce applications.

In order to be able to put forward an effective m-commerce privacy security assessment method, this paper will collate relevant risk factors, establish a multilevel and multiangle assessment model, which constructs the hierarchy analysis model of privacy risk, uses information entropy to describe privacy risks, simulates and analyses a real risk environment of m-commerce application based on Markov chain, and realizes the effective assessment of the privacy security of m-commerce users. The privacy security risk assessment method proposed in this paper aim to provide a comprehensive method for the accurate and quantitative evaluation of privacy disclosure risk in real risk environment of m-commerce application.

3. Method Development for Risk Assessment Based on Information Entropy and Markov Chain

For the purpose of risk assessment, this paper proposes to integrate information entropy and Markov chain into the privacy risk assessment of m-commerce users; the framework of our work is shown in Figure 1.

As shown in Figure 1, our proposed assessment method is developed to integrate the works on the three parts.

3.1. Development of Risk Measurement Method Based on Information Entropy. Information entropy was proposed by Shannon in 1948. In Shannon's theory, information entropy

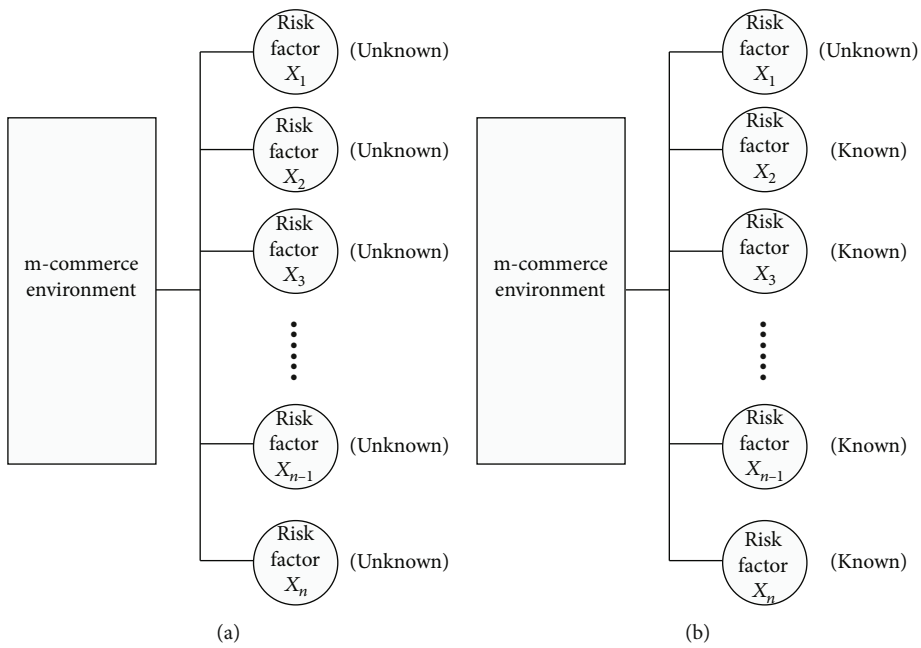


Figure 2: Comparison of two extreme risk factors in m-commerce environment.

is measured by the well-known formula $H(X) = -\sum_{i=1}^{n} P(X_i) \log_2 P(X_i)$, where X_i is the information source variable and $P(X_i)$ is the probability of the information source. In information theory, information entropy is used to represent the amount of information content and quantify the uncertainty of things.

Privacy security is not so objective to be easily measured. However, with the method of information entropy, it can be described from the perspective between known to unknown, the two opposite extremes. That is, the privacy risks of users are described with the characteristics of information entropy uncertainty, as shown in Figure 2.

Figure 2(a) shows the mobile business environment with n unknown risks X_i, i.e., $X = \{X_1, X_2, \cdots, X_n\}$; according to the information entropy theory, its entropy value $H(X)$ will reach its maximum, $H(X) = \log_2 n$, when all the risks occur with the same probability, that is $P(X_1) = P(X_2) =, \cdots, = P(X_n)$. This idea also suggests that the higher the user privacy risk uncertainty, the lower the controllability of the risks, and the lower the security.

Figure 2(b) shows the opposite case, when there is only one unknown risk in the m-commerce environment and the other risks are controllable, the entropy value $H(X)$ will reach its minimum according to the information entropy theory, indicating that the lower the privacy security risk uncertainty of the application, the higher the security, namely, the risk is substantially controllable.

3.2. Simulation of Risk Environment Based on Markov Chain. Markov chain [30, 31] is a discrete time random process of continuous transition from one state to another in the finite state space. It can describe the state space of the change of state of things and calculate the probability of occurrence of each random state of things by establishing Markov chain transfer matrix.

In addition to effective risk measurement method, the user privacy disclosure risk of m-commerce still needs to be assessed, and the random state in the practical application shall be analyzed, so as to ensure the validity of the assessment results. Therefore, the complex environment of user privacy disclosure risk in m-commerce is described in line with Markov chain, to achieve effective assessment of the user privacy security based on the practical conditions.

Assuming that there are n risk factors X_i in an m-commerce environment, according to Markov chain, this complex risk environment can be described as the following matrix taking into account the mutual influence between every two factors:

$$R = \begin{bmatrix} X_{11}X_{12} & & \cdots & X_{1n} \\ X_{21} & X_{22} & & X_{2n} \\ \vdots & & \ddots & \vdots \\ X_{n1} & & X_{n2} & \cdots & X_{nn} \end{bmatrix}. \quad (1)$$

Matrix R is an m-commerce privacy risk matrix, where the elements X_{ii} on the diagonal line represent the separate occurrence of risk factors X_i, and X_{ij} represent simultaneous occurrence of risk factors X_i and X_j in the actual application process. The matrix R represents the complex privacy risk environment of m-commerce users by mathematical method, which provides a guarantee for the simulation analysis of this paper.

3.3. Construction of Risk Hierarchy. In the above discussions, we outline a method for descripting the privacy risk based

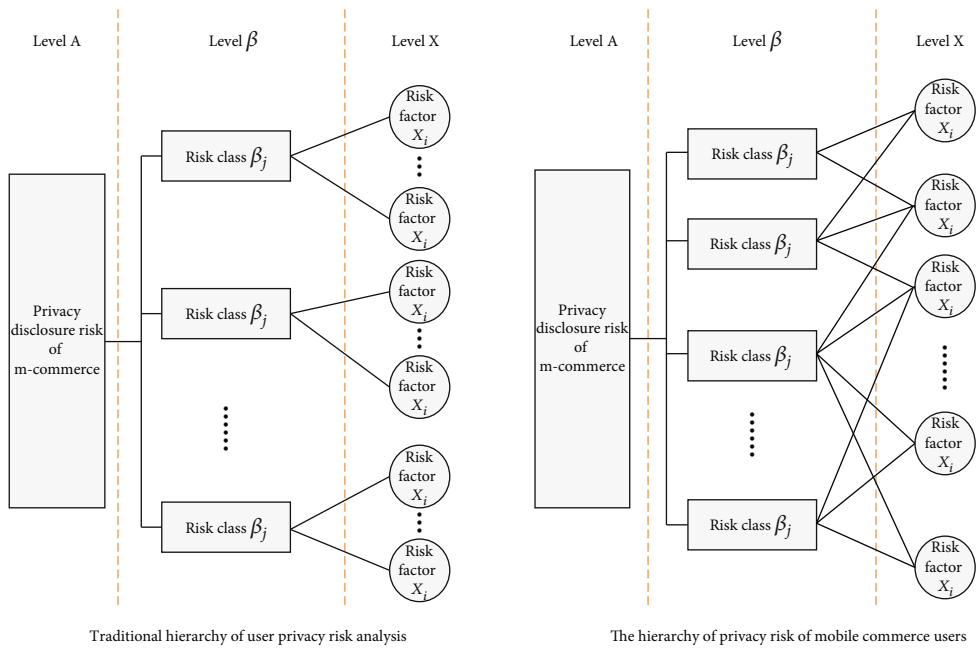


Figure 3: The hierarchy of privacy risk of m-commerce users.

Table 2: Risk categories β_1 and β_2 and the risk factors they contain.

Risk class	Risk factors included risk class
β_1	X_1, X_2, X_3
β_2	X_3, X_4

on information entropy, and we develop a Markov matrix to simulate the complexity of risk environment for m-commerce. In our framework, it is still necessary to further establish a risk hierarchy to allow for multidimensional and multilevel simulation analysis of m-commerce user privacy risk, which is shown in Figure 3.

This hierarchy consists of three levels, target level A, risk class level β, and risk factor level X. Each risk class β_j includes multiple risk factors X_i. Different from the traditional user privacy risk analysis, our proposed analyzing framework presents a cross relationship between risk factors and risk class, which is more consistent with the real risk environment of m-commerce.

3.4. Development of the Proposed Assessment Method. A bottom-up process is used to the hierarchy in our method. In the following discussions, we use $P(X_i)$ to represent the probability of occurrence of risk X_i at level 3, and normalization process is carried out based on the classified categories, to calculate the probabilities $P(X_{ij})$ $i, j = 1, 2, \cdots, n$ of risk occurrence under different categories, which are substituted into the matrix R to further obtain the state transition matrix $P(R)$ of the m-commerce privacy.

The calculation process is as shown in the following example: it is assumed that there are two risk classes, namely, β_1 and β_2, which include risk factors as shown in Table 2.

As shown in Table 1, class β_1 includes particular risk factor X_1, class β_2 includes particular risk factor X_4, while risk factor X_3 is included in both β_1 and β_2, then their transition state matrix can be derived through calculation.

$$P(R) = \begin{bmatrix} P(\beta_{11}) & P(\beta_{12}) \\ P(\beta_{21}) & P(\beta_{22}) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sum_{i=1}^{3} P(X_i)} P(X_1) + P(X_2) & \frac{1}{\sum_{i=1}^{3} P(X_i)} P(X_3) \\ \frac{1}{\sum_{i=3}^{4} P(\alpha_i)} P(X_3) & \frac{1}{\sum_{i=3}^{4} P(\alpha_i)} P(X_4) \end{bmatrix}. \tag{2}$$

Similarly, according to formula (2), it is assumed that there are m risk classes β_i and n risk factors X_i in an m-commerce, then the privacy risk transfer matrix $P(R)$ for this m-commerce application can be derived based on the classified classes.

$$P(R) = \begin{bmatrix} P(X_{11}) & P(X_{12}) & \cdots & P(X_{1m}) \\ P(X_{21}) & P(X_{22}) & & P(X_{2m}) \\ \vdots & & \ddots & \vdots \\ P(X_{m1}) & P(X_{m2}) & \cdots & P(X_{mm}) \end{bmatrix}. \tag{3}$$

It is assumed that in the long utilization, the steady-state probability of class β_i is $\widehat{P}(\beta_i)$, $\mathrm{i} = 1, 2, \cdots, m$. It is a

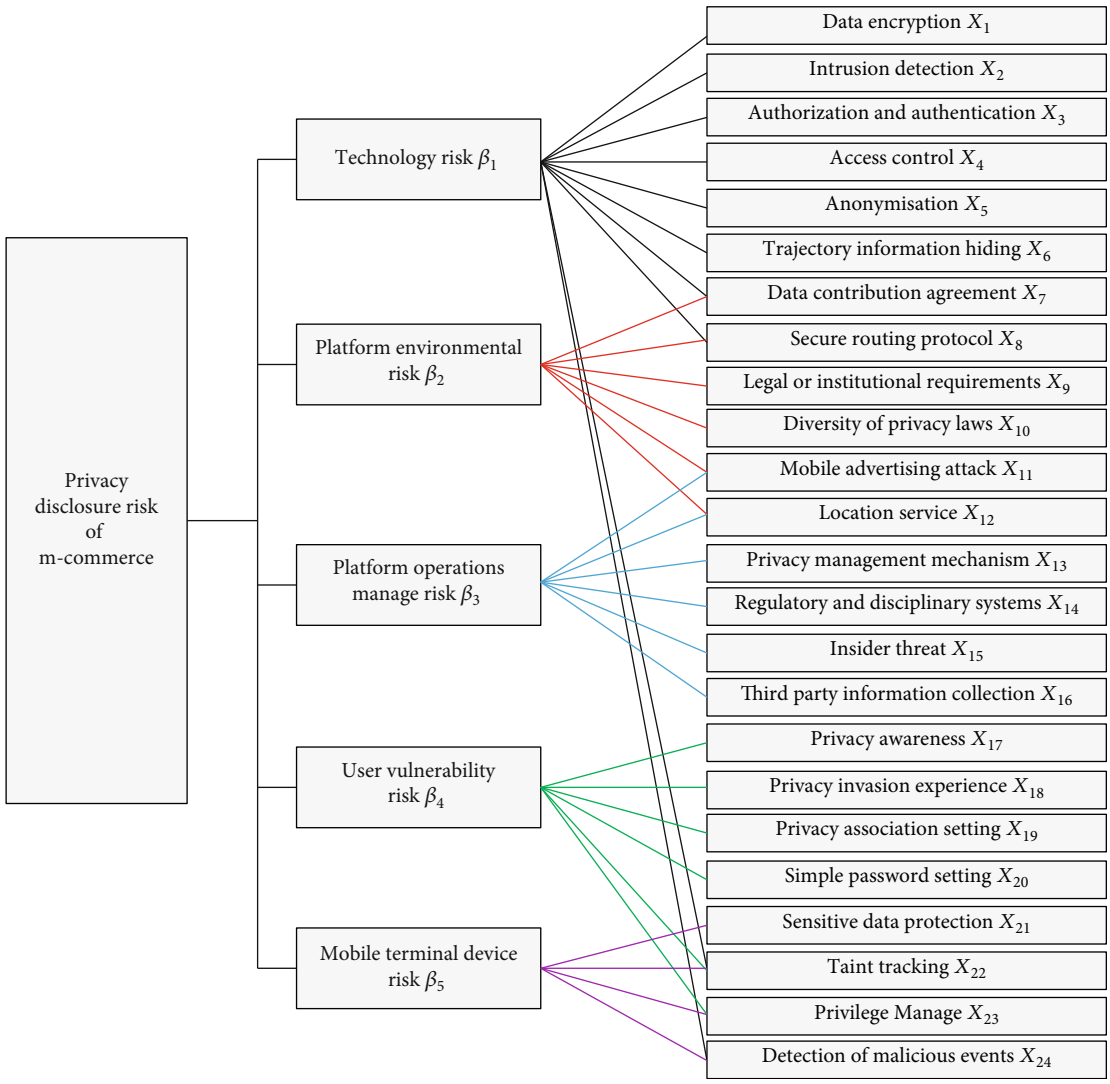


Figure 4: Hierarchical attribute model of privacy disclosure risk in m-commerce.

possible probability of a certain risk class in the long stable utilization and a stable probability calculated by the Markov method. According to this method, the relation between $\widehat{P}(\beta_i)$ and sate transition matrix $P(R)$ satisfies the following equation:

$$\begin{cases} \widehat{P}(\beta_1) = P(X_{11})\widehat{P}(\beta_1) + \mathrm{P}(X_{12})\widehat{P}(\beta_2) + \cdots + \mathrm{P}(X_{1m})\widehat{P}(\beta_m) \\ \widehat{P}(\beta_2) = P(X_{21})\widehat{P}(\beta_1) + \mathrm{P}(X_{22})\widehat{P}(\beta_2) + \cdots + \mathrm{P}(X_{2m})\widehat{P}(\beta_m) \\ \widehat{P}(\beta_3) = P(X_{31})\widehat{P}(\beta_1) + \mathrm{P}(X_{32})\widehat{P}(\beta_2) + \cdots + \mathrm{P}(X_{3m})\widehat{P}(\beta_m) \\ \vdots \\ \widehat{P}(\beta_{\mathrm{m}}) = P(X_{m1})\widehat{P}(\beta_1) + \mathrm{P}(X_{m2})\widehat{P}(\beta_2) + \cdots + \mathrm{P}(X_{mm})\widehat{P}(\beta_m) \\ 1 = \widehat{P}(\beta_1) + \widehat{P}(\beta_2) + \cdots + \widehat{P}(\beta_m) \end{cases}. \tag{4}$$

The occurrence probability of various risks $\widehat{P}(\beta_i) = \{\widehat{P}(\beta_1), \widehat{P}(\beta_2), \cdots, \widehat{P}(\beta_m)\}$, $\sum_{i=1}^{m}\widehat{P}(\beta_i) = 1$ in the longstable utilization of the m-commerce application can be derived by solving the Equation (3).

Table 3: The level of probability of risk factors occurrence.

Level	Definition and description
(8, 10)	This factor has a great risk and a direct threat to the user's privacy
(6, 8)	This risk has a high probability of occurrence and exists in most m-commerce environments
(4, 6)	This risk is a common risk, which exists in some m-commerce
(2, 4)	This risk exists and only occurs when special conditions are met
(0, 2)	This factor has high security and hardly causes user privacy risk

Therefore, the privacy risk assessment results H of the entire m-commerce environment can be calculated by

Table 4: Scoring results of probability of occurrence of underlying risk factors.

Company	Risk factor x_i	Level	$P(x_i)$	Risk factor x_i	Level	$P(x_i)$	Risk factor x_i	Level	$P(x_i)$
A	X_1	2	1.887%	X_9	2	1.887%	X_{17}	9	8.491%
	X_2	2	1.887%	X_{10}	2	1.887%	X_{18}	9	8.491%
	X_3	3	2.830%	X_{11}	7	6.604%	X_{19}	7	6.604%
	X_4	3	2.830%	X_{12}	7	6.604%	X_{20}	6	5.660%
	X_5	1	0.943%	X_{13}	2	1.887%	X_{21}	5	4.717%
	X_6	1	0.943%	X_{14}	3	2.830%	X_{22}	4	3.774%
	X_7	6	5.660%	X_{15}	2	1.887%	X_{23}	6	5.660%
	X_8	6	5.660%	X_{16}	7	6.604%	X_{24}	4	3.774%
B	X_1	2	2.000%	X_9	2	2.000%	X_{17}	9	9.000%
	X_2	7	7.000%	X_{10}	2	2.000%	X_{18}	8	8.000%
	X_3	3	3.000%	X_{11}	3	3.000%	X_{19}	9	9.000%
	X_4	2	2.000%	X_{12}	2	2.000%	X_{20}	1	1.000%
	X_5	2	2.000%	X_{13}	5	5.000%	X_{21}	5	5.000%
	X_6	2	2.000%	X_{14}	3	3.000%	X_{22}	4	4.000%
	X_7	4	4.000%	X_{15}	2	2.000%	X_{23}	7	7.000%
	X_8	4	4.000%	X_{16}	8	8.000%	X_{24}	4	4.000%
C	X_1	1	0.901%	X_9	2	1.802%	X_{17}	9	8.108%
	X_2	1	0.901%	X_{10}	2	1.802%	X_{18}	9	8.108%
	X_3	3	2.703%	X_{11}	2	1.802%	X_{19}	9	8.108%
	X_4	2	1.802%	X_{12}	9	8.108%	X_{20}	2	1.802%
	X_5	1	0.901%	X_{13}	8	7.207%	X_{21}	3	2.703%
	X_6	9	8.108%	X_{14}	3	2.703%	X_{22}	4	3.604%
	X_7	8	7.207%	X_{15}	2	1.802%	X_{23}	8	7.207%
	X_8	3	2.703%	X_{16}	7	6.306%	X_{24}	4	3.604%

substituting $\widehat{P}(\beta_i)$ into the following information entropy formula (5):

$$H = -\sum_{i=1}^{m} \widehat{P}(\beta_i) \log_2 \widehat{P}(\beta_i), \tag{5}$$

where H represents the entropy value for privacy security of the m-commerce users, and the greater its value, the lower the privacy security of the m-commerce. The entropy value of the risk class β_i can be derived by normalizing the occurrence probability of risk factors included in such class following the information entropy calculation method, and the greater this value, the lower the privacy security of this risk class.

4. Integration of the Assessment Method

4.1. Risk Attribute Model for Privacy Disclosure of m-commerce Users. According to the assessment method proposed in Section 3, 24 risk evaluation indicators for privacy information disclosure of m-commerce users are selected, and these indicators are divided into 5 classes, i.e., technology risk, platform environmental risk, platform operation manage risk, user vulnerability risk, and mobile terminal device risk. According to the hierarchical structure in Figure 3, a hierarchical attribute model for privacy disclosure risk is built, as shown in Figure 4.

4.2. Measurement and Assessment of Privacy Disclosure Risks. Based on the m-commerce user privacy risk attribute model in Figure 4 and in accordance with the assessment method proposed herein, the detailed calculation process is as follows:

Step 1. Table 3 "the level of probability of risk factors occurrence" is prepared, and the occurrence probability level of the lowest-level risk factors is obtained through scoring by experts, and the values of $P(X_i)$ obtained through normalization processing.

Step 2. Based on the hierarchical structure in Figure 4 and according to Markov chain, use Equation (2) to calculate the state transition matrix $P(R)$.

Table 5: The steady-state probability of risk class.

Company	β_i	$\widehat{P}(\beta_i)$	Company	β_i	$\widehat{P}(\beta_i)$	Company	β_i	$\widehat{P}(\beta_i)$
	β_1	0.159		β_1	0.119		β_1	0.135
	β_2	0.181		β_2	0.163		β_2	0.172
A	β_3	0.190	B	β_3	0.203	C	β_3	0.192
	β_4	0.247		β_4	0.277		β_4	0.266
	β_5	0.228		β_5	0.247		β_5	0.242

Table 6: Comparison of evaluation results of three companies.

Company	Risk entropy H
A	2.307
B	2.270
C	2.288

Step 3. Use Equation (4) to calculate the stability probability $\widehat{P}(\beta_i)$ of various risks.

Step 4. Use formula (5) to calculate H, so as to evaluate the privacy security of the entire m-commerce environment.

Step 5. Normalize the probability of occurrence of these risk factors to obtain their weight coefficients $P(X_j, \beta_i)$ in different risk classes. Then, calculate various risk entropy $H(\beta_i)$ in combination with the information entropy formula with the following.

$$H(\beta_i) = \frac{-\sum_{j=1}^{m} P(X_j, \beta_i) \log_2 P(X_j, \beta_i)}{\log_2 m}, \tag{6}$$

where m is the number of risk factors included in risk class β_i. The larger this value is, the more difficult it is to control such risks, and the greater the privacy security risk will be.

5. Case Study

5.1. Assessment Process. In order to verify the feasibility of the proposed method, three companies with different nature in m-commerce applications background are selected and assessed from bottom to top in details, where company A provides food delivery m-commerce service, company B provides financial m-commerce service, and company B provides map navigation service. The three applications all carry the users' privacy data like information of finance, identity, location, and device. The assessment is specifically carried out for these three companies as follows:

Step 1. First of all, the bottom risk factors x_i of three m-commerce applications are scored by a panel of 10 experts with AHP [32] method according to the definitions in Table 3. After the scoring is completed, the scores of 10 experts are summed up, averaged to obtain their level, and the level is further normalized to obtain the value of $P(x_i)$, and the results are shown in Table 4.

Table 7: Normalization results of risk factors contained in each risk class.

Risk class β_i	Risk factor x_j contained in β_i
β_1	$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_{22}, x_{24}\}$
β_2	$\{x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$
β_3	$\{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}\}$
β_4	$\{x_{17}, x_{18}, x_{19}, x_{20}, x_{22}, x_{23}\}$
β_5	$\{x_{21}, x_{22}, x_{23}, x_{24}\}$

Step 2. Based on the hierarchical structure in Figure 4, the results of Table 4 are substituted into formula (2), to obtain the following state transition matrices $P^{A}(R)$, $P^{B}(R)$, and $P^{C}(R)$ for privacy disclosure risk of m-commerce users of the three companies.

$$P^{A}(R) = \begin{bmatrix} 0.375 & 0.375 & 0.000 & 0.000 & 0.250 & & \\ 0.400 & 0.133 & 0.467 & 0.000 & 0.000 & & \\ 0.000 & 0.500 & 0.500 & 0.000 & 0.000 & & \\ 0.000 & 0.000 & 0.000 & 0.756 & 0.244 & & \\ 0.211 & & 0.000 & & 0.000 & 0.526 & 0.263 \end{bmatrix},$$
$$P^{B}(R) = \begin{bmatrix} 0.529 & 0.235 & 0.000 & 0.000 & 0.235 & & \\ 0.471 & 0.235 & 0.294 & 0.000 & 0.000 & & \\ 0.000 & 0.217 & 0.783 & 0.000 & 0.000 & & \\ 0.000 & 0.000 & 0.000 & 0.711 & 0.289 & & \\ 0.200 & & 0.000 & & 0.000 & 0.550 & 0.250 \end{bmatrix},$$
$$P^{C}(R) = \begin{bmatrix} 0.472 & 0.306 & 0.000 & 0.000 & 0.222 & & \\ 0.423 & 0.154 & 0.423 & 0.000 & 0.000 & & \\ 0.000 & 0.355 & 0.645 & 0.000 & 0.000 & & \\ 0.000 & 0.000 & 0.000 & 0.707 & 0.293 & & \\ 0.211 & & 0.000 & & 0.000 & 0.632 & 0.158 \end{bmatrix}. \tag{7}$$

Step 3. The data in the above transition matrices are substituted into formula (4) to calculate the steady-state probability of various risks, as shown in Table 5.

Step 4. The calculated results of Table 5 are substituted into formula (5) to obtain the user privacy security evaluation results of three companies' m-commerce applications, as shown in Table 6.

Step 5. The risk factors included in different risk classes are further normalized. The known risk classes and the contained risk factors are shown in Table 7.

Based on the division of Table 7, the level-2 risk classes of three different m-commerce companies are evaluated in this paper, and the calculated entropy values of various risks are shown in Figure 5.

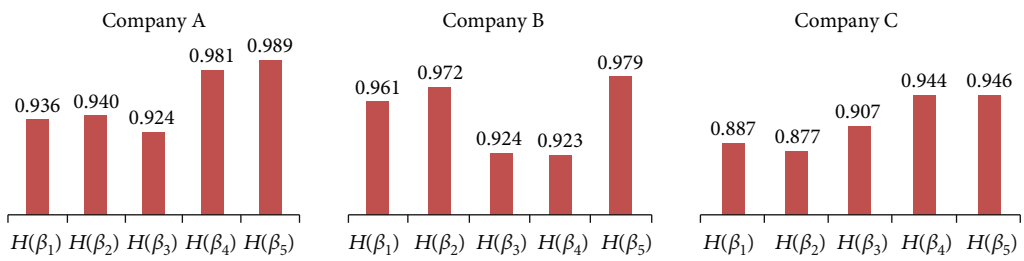


Figure 5: Comparison of entropy values of various risk class.

5.2. Analysis of Assessment Results. According to the results of the assessment and the proposed risk hierarchy, the analysis is carried out level by level:

5.2.1. Analysis of Top-Level Evaluation Results. The comparison of Table 6 shows that $H(\mathrm{A}) > H(\mathrm{C}) > H(\mathrm{B})$, indicating that the food delivery m-commerce application of company A has higher privacy risk compared with the other two applications; on the contrary, company B's financial m-commerce application enjoys the highest privacy security.

However, the privacy security evaluation results of the three companies are not very different in data size, indicating that on the whole, the three companies have similar privacy security performance and certain privacy and security factors.

5.2.2. Analysis of Middle-Level Evaluation Results

(1) Comparative Analysis of Steady-State Probability Results of Risk Classes. It is found in Table 5 that $\widehat{P}(\beta_4) > \widehat{P}(\beta_5) > \widehat{P}(\beta_3) > \widehat{P}(\beta_2) > \widehat{P}(\beta_1)$. This result shows that the three companies share one marked characteristic, namely, the value $\widehat{P}(\beta_4)$ is the greatest, indicating that compared with other risk classes, user vulnerability risk β_4 is most likely to occur in the long utilization of m-commerce application. Secondly, the value of $\widehat{P}(\beta_5)$ is great, which suggests that while user vulnerability risk can easily arise, the terminal device often causes security problems. On the contrary, the value of the technical risk $\widehat{P}(\beta_1)$ is the smallest, which indicates that technology risk is not the main cause of the user's privacy information security problem, and compared with other risk classes, it is not likely that the privacy security problems are caused by technology risk.

Thus, it can be seen that when m-commerce users disclose personal information in pursuit of personalized services provided by the m-commerce platform, there are mainly problems such as low awareness of privacy risks, numerous privacy association settings, insufficient experience in privacy invasion, and simple password setting, etc. The above situation poses a great threat to user privacy. Users should strengthen their awareness of privacy protection and improve their ability to deal with risks. While enjoying the convenience brought by m-commerce, users should also understand the risks and avoid excessive disclosure of their private information.

(2) Analysis of Comparison Results of Entropy Value of Risk Classes. Figure 5 shows that the $H(\beta_5)$ values of the three companies are the greatest, indicating that it is most difficult to control the mobile terminal device risk.

Moreover, the evaluation results show high $H(\beta_4)$ of company A and company C, indicating that when utilizing the m-commerce applications of these two companies, the users could hardly control their own privacy risk, giving rise to privacy security issues in these applications (take-away catering, map navigation). By contrast, the value of financial application $H(\beta_4)$ is low, which suggests that the users' behavior and operation are strictly regulated in such application, and its user risk is easier to control compared with other applications. This comparison shows that platform environment risk $H(\beta_2)$ should be mainly blamed for the leakage of such application privacy information.

5.2.3. Analysis of Bottom-Level Evaluation Results. The above comparison shows that the user vulnerability risk β_4 has the highest probability of occurrence. There is an observation of the bottom factors of such risk class that the security problems of m-commerce applications are mainly caused by the users' weak privacy risk awareness x_{17}, excessive privacy association settings x_{18}, the lack of privacy invasion experience x_{19}, and simple password setting x_{20} and so on.

In the financial application, the platform environment risk β_2 has the greater probability of occurrence, it is affected by x_7, x_8, x_9, and so on. This result shows that the privacy security problem is mainly caused by the platform environment risk factors such as the data sharing agreement with the users x_7, the security routing agreement x_8, the formulation of privacy law x_9, and so on.

5.3. Suggestions and Remarks. It is well known that it is not feasible to only improve the privacy security of m-commerce users by the use of technique tools. The current risk problems mainly arise from the weak privacy security awareness of users, and the security issues of m-commerce applications will not be effectively solved until the users are more aware of and better understand the privacy security issues. For this purpose, the operator should more diligently remind the users on privacy security in the utilization of m-commerce applications, standardize their relevant operation as much as possible, and urge them to take security protection measures. On the other hand, for some financial applications, more explicit confidentiality agreement shall be

signed with the users, the access to the users' permission shall be reduced, relevant responsibilities shall be clarified, and guarantee the information security of the users through laws and regulations.

6. Conclusions

In this paper, the risk factors of user privacy disclosure in m-commerce are reviewed, the magnitude of risks is measured based on information entropy, to provide effective data support for risk assessment. We have detailed discussed the complexity of user privacy risk in the real environment, and a complete assessment model for user privacy disclosure risks is established, and reasonable risk measurement and assessment methods are proposed based on Markov chain. In addition, a detailed comparative analysis is carried out based on the actual application that can provide practical reference for the protection of the privacy security of m-commerce users, and enrich and improve the relevant research theory of user privacy security. In the future research, with the update of the m-commerce application service, it is necessary to keep track of the latest research theories and further improve the attribute model of the user privacy risk. Moreover, the risks can be divided in line with actual application into more classes, which can be selected based on relevant risk factors, to realize more accurate assessment and research on user privacy security.

Data Availability

The expert scoring data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation Project (No. 71462036), Yunnan Basic Applied Research Program (No. 2017FD130), the Scientific Research Foundation of Yunnan Education Department (No. 2020J0377, No. 2020J0392, and No. 2020J0378), and School-level Project of Yunnan University of Finance and Economics (No. 2017D29, No. 2018B07, and No. 2018B08).

References

[1] C. Cao and X. Zhu, "Strong anonymous mobile payment against curious third-party provider," *Electronic Commerce Research*, vol. 19, no. 3, pp. 501–520, 2019.

[2] Y. Lei, "The top 10 data breaches of 2019," *Computer Networks*, vol. 46, no. 2, pp. 46-47, 2020.

[3] Y. Z. Xu, J. L. Zhang, Y. Hua, and L. Y. Wang, "Dynamic Credit Risk Evaluation Method for E-Commerce Sellers Based on a Hybrid Artificial Intelligence Model," *Sustainability*, vol. 11, no. 19, 2019.

[4] Y. Zhang, R. Wang, M. S. Hossain, M. F. Alhamid, and M. Guizani, "Heterogeneous Information Network-Based Content Caching in the Internet of Vehicles," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 10216–10226, 2019.

[5] Y. Wu, G. Feng, N. Wang, and H. Liang, "Game of information security investment: Impact of attack types and network vulnerability," *Expert Systems with Applications*, vol. 42, no. 15-16, pp. 6132–6146, 2015.

[6] F. Shirazi and A. Iqbal, "Community clouds within M-commerce: a privacy by design perspective," *Journal of Cloud Computing*, vol. 6, no. 1, 2017.

[7] E. Aghasian, S. Garg, and J. Montgomery, "A Privacy-Enhanced Friending Approach for Users on Multiple Online Social Networks," *Computers*, vol. 7, no. 3, article 7030042, 2018.

[8] Y. Zhang, Y. Qian, M. S. H. Di Wu, A. Ghoneim, and M. Chen, "Emotion-Aware Multimedia Systems Security," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 617–624, 2019.

[9] Y. Ashibani and Q. H. Mahmoud, "Cyber physical systems security: Analysis, challenges and solutions," *Computers and Security*, vol. 68, pp. 81–97, 2017.

[10] H. Zhu, C. X. J. Ou, W. J. A. M. van den Heuvel, and H. Liu, "Privacy calculus and its utility for personalization services in e-commerce: An analysis of consumer decision-making," *Information and Management*, vol. 54, no. 4, pp. 427–437, 2017.

[11] M. Fodor and A. Brem, "Do privacy concerns matter for Millennials? Results from an empirical analysis of location-based services adoption in Germany," *Computers in Human Behavior*, vol. 53, pp. 344–353, 2015.

[12] V. M. Wottrich, E. A. van Reijmersdal, and E. G. Smit, "The privacy trade-off for mobile app downloads: The roles of app value, intrusiveness, and privacy concerns," *Decision Support Systems*, vol. 106, pp. 44–52, 2018.

[13] A. Gutierrez, S. O'Leary, N. P. Rana, Y. K. Dwivedi, and T. Calle, "Using privacy calculus theory to explore entrepreneurial directions in mobile location-based advertising: Identifying intrusiveness as the critical risk factor," *Computers in Human Behavior*, vol. 95, pp. 295–306, 2019.

[14] G. Ampong, A. Mensah, A. Adu, J. Addae, O. Omoregie, and K. Ofori, "Examining Self-Disclosure on Social Networking Sites: A Flow Theory and Privacy Perspective," *Behavioral Sciences*, vol. 8, no. 6, pp. 58–75, 2018.

[15] G. Zhu, M. N. Feng, Y. Chen, and J. Y. Yang, "Research on Fuzzy Evaluation of Privacy Risk for Social Network in Big Data Environment," *Information Science*, vol. 34, no. 9, pp. 94–98, 2016.

[16] B. Tian, Y. S. Zheng, P. Y. Liu, and C. H. Li, "The evaluation index and empirical study on risk of privacy information disclosure of mobile APP users," *Library and Information Services*, vol. 62, no. 19, pp. 101–110, 2018.

[17] M. M. Xiang, X. W. Wang, R. N. Jia, and L. Wang, "Research on the Risk Evaluation of Consumers' Privacy Information Disclosure in Mobile Commerce," *Library and information service*, vol. 62, no. 18, pp. 34–44, 2018.

[18] Y. Nan, Z. Yang, M. Yang et al., "Identifying User-Input Privacy in Mobile Applications at a Large Scale," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 3, pp. 647–661, 2017.

[19] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge Intelligence in the Cognitive Internet of Things: Improving Sensitivity and Interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58–64, 2019.

[20] H. Li, B. Wang, W. Zhang, Q. Tang, and Y. L. Zhang, "X-Decaf : Detection of Cache File Leaks in Android Social Apps," *Journal of Electronics & Information Technology*, vol. 39, no. 1, pp. 66–74, 2017.

[21] Y. A. Tan, Y. Xue, C. Liang et al., "A root privilege management scheme with revocable authorization for Android devices," *Journal of Network and Computer Applications*, vol. 107, pp. 69–82, 2018.

[22] A. Ruiz-Heras, P. García-Teodoro, and L. Sánchez-Casado, "ADroid: anomaly-based detection of malicious events in Android platforms," *International Journal of Information Security*, vol. 16, no. 4, pp. 371–384, 2017.

[23] T. L. Gao, T. Li, R. Jiang, M. Yang, and R. Zhu, "Research on cloud service security measurement based on information entropy," *International Journal of Network Security*, vol. 21, no. 6, pp. 1003–1013, 2019.

[24] M. Yang, R. Jiang, T. L. Gao, W. Y. Xie, and J. Wang, "Research on cloud computing security risk assessment based on information entropy and Markov chain," *International Journal of Network Security*, vol. 20, no. 4, pp. 664–673, 2018.

[25] G. Tilei, L. Tong, Y. Ming, and J. Rong, "Research on a Trustworthiness Measurement Method of Cloud Service Construction Processes Based on Information Entropy," *Entropy*, vol. 21, no. 5, 2019.

[26] J. Wang, J. Liu, and H. Zhang, "Access Control Based Resource Allocation in Cloud Computing Environment," *International Journal of Network Security*, vol. 19, no. 2, pp. 236–243, 2017.

[27] Y. C. Wei, W. C. Wu, G. H. Lai, and Y. C. Chu, "pISRA: privacy considered information security risk assessment model," *The Journal of Supercomputing*, vol. 76, no. 3, pp. 1468–1481, 2020.

[28] M. C. Oetzel and S. Spiekermann, "A systematic methodology for privacy impact assessments: a design science approach," *European Journal of Information Systems*, vol. 23, no. 2, pp. 126–150, 2019.

[29] N.-W. Lo, K.-H. Yeh, and C.-Y. Fan, "Leakage Detection and Risk Assessment on Privacy for Android Applications: LRPdroid," *IEEE Systems Journal*, vol. 10, no. 4, pp. 1361–1369, 2016.

[30] W. J. Stewart, *Introduction to the numerical solutions of Markov chains*, USA:Princeton University Press, Princeton, 1994.

[31] W. J. Stewart, *Probability, Markov Chains, Queues, and Simulation: the Mathematical Basis of Performance Modeling. Princeton*, Princeton University Press, Princeton,USA, 2009.

[32] M. Yang, T. B. Li, R. Jiang, T. L. Gao, and J. Wang, "Research on Model of Big Data Usability and Mining Strategy Based on AHP," *Computer Technology and Development*, vol. 28, no. 5, pp. 51–58, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8816228, 12 pages
https://doi.org/10.1155/2020/8816228

Research Article

An Efficient Algorithm for Extracting High-Utility Hierarchical Sequential Patterns

Chunkai Zhang, Zilin Du, and Yiwen Zu

School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

Correspondence should be addressed to Chunkai Zhang; ckzhang@hit.edu.cn

Received 18 March 2020; Revised 8 June 2020; Accepted 23 June 2020; Published 6 July 2020

Academic Editor: Huimin Lu

High-utility sequential pattern mining (HUSPM) is an emerging topic in data mining, where utility is used to measure the importance or weight of a sequence. However, the underlying informative knowledge of hierarchical relation between different items is ignored in HUSPM, which makes HUSPM unable to extract more interesting patterns. In this paper, we incorporate the hierarchical relation of items into HUSPM and propose a two-phase algorithm MHUH, the first algorithm for high-utility hierarchical sequential pattern mining (HUHSPM). In the first phase named Extension, we use the existing algorithm FHUSpan which we proposed earlier to efficiently mine the general high-utility sequences (g-sequences); in the second phase named Replacement, we mine the special high-utility sequences with the hierarchical relation (s-sequences) as high-utility hierarchical sequential patterns from g-sequences. For further improvements of efficiency, MHUH takes several strategies such as Reduction, FGS, and PBS and a novel upper bounder TSWU, which will be able to greatly reduce the search space. Substantial experiments were conducted on both real and synthetic datasets to assess the performance of the two-phase algorithm MHUH in terms of runtime, number of patterns, and scalability. Conclusion can be drawn from the experiment that MHUH extracts more interesting patterns with underlying informative knowledge efficiently in HUHSPM.

1. Introduction

Sequential pattern mining (SPM) [1–3] is an interesting and critical research area in data mining. According to the problem definition [4], a large database of customer transactions has three fields, i.e., customer-id, transaction-time, and the items bought. Each transaction corresponds to an itemset, and all the transactions from a customer are ordered by increasing transaction-time to form a sequence called customer sequence. The support of a sequence is the number of customer sequences that contains it. If the support of a sequence is larger than a user-specified minimum support, we call it a frequent sequence. The sequential pattern mining algorithm will discover the frequent sequences called sequential patterns among all sequences. In a word, the purpose of sequential pattern mining is to discover all frequent sequences as sequential patterns, which reflect the potential connections within items, from a sequence database under the given minimum support. An example of such a sequential pattern is that customers typically buy a phone, then a phone shell, and then a phone charger. Customers who buy some other commodities in between also support this sequential pattern. In the past decades, many algorithms [1, 5] have been proposed for sequential pattern mining, which makes it be widely applied in many realistic scenarios (e.g., consumer behavior analysis [6] and web usage mining [7]). However, sequential pattern mining has two apparent limitations.

Firstly, frequency does not fully reveal the importance (i.e., interest) in many situations [8–12]. In fact, many rare but important patterns may be missed under the frequency-based framework. For example, in retail selling, a phone usually brings more profit than selling a bottle of milk, while the quantity of phones sold is much lower than that of milk [9], and the decision-maker tends to emphasize the sequences consisting of high-profit commodities, instead of those frequent commodity sequences. This issue leads to the emergence of high-utility sequential pattern mining (HUSPM) [8, 12–15]. To represent the relative importance of patterns, each item in the database is associated with a value called external utility (e.g., indicating the unit profit of the item

purchased by a customer). In addition, each occurrence of the item is associated with a quantity called internal utility (e.g., indicating the number of units of the item purchased by a customer in a transaction). The utility of a sequence is computed by applying a utility function on all sequences in the database where it appears. The task of high-utility sequential pattern mining is to discover all high-utility sequential patterns (HUSPs, the sequences with high utility) from the quantitative sequence database with a predefined minimum utility threshold. Many high-utility sequential pattern mining algorithms have been proposed in the past decades [13, 16–20], and high-utility sequential patterns can be extracted more efficiently with a series of novel data structures and pruning strategies proposed. In addition, high-utility sequential pattern mining has many practical applications including web log data [21], mobile commerce environments [22], and gene regulation data.

Secondly, in sequential pattern mining, the hierarchical relation (e.g., product relation and semantic relation) between different items is ignored, so some underlying knowledge may be missed. In general, the individual items of the input sequences are naturally arranged in a hierarchy [23]. For example, suppose both the sequence S_1 : <phone, mobile power pack > and the sequence S_2 : <phone, bluetooth headset > are infrequent, then it seems that there is no association between the three commodities. However, we may find that the sequence S_3 : <phone, phone accessory > is frequent from the perspective of product hierarchy, indicating that customers usually buy a phone first, then buy a phone accessory (including "mobile power pack" and "bluetooth headset"). That is to say, products in sequences of customer transactions can be arranged in a product hierarchy, where mobile power pack and bluetooth headset can generalize phone accessory. Another example is that the individual word in a text can form a semantic hierarchy. The words drives and driven can generalize to their common lemma drive, which in turn generalize to their respective part-of-speech tag verb. The concept of hierarchy (is a taxonomy) provides the deciders with a different perspective to analyze sequential patterns. More informative patterns can be extracted through the hierarchy-based methodology. Besides, although the information reveled from the hierarchical perspective may be relatively fuzzy, it reduces the loss of underlying knowledge to a certain extent. Particularly, the hierarchical relation between different items is sometimes inherent to the application (e.g., hierarchies of directories or web pages) or they are constructed in a manual or automatic way (e.g., product relation) [23]. Figure 1 shows a simple example of a taxonomy of biology in the real application. Sequential pattern mining with hierarchical relation can be traced back to the article [6] where the hierarchy management was incorporated into sequential pattern mining, and the GSP algorithm was proposed to extract sequential patterns according to different levels of hierarchy. Later, sequential pattern mining with hierarchical relation has been studied extensively in the literature [24, 25]. Efficient algorithms were proposed in a wide range of real-world applications, such as customer behavior analysis [6, 26] and information extraction [27].

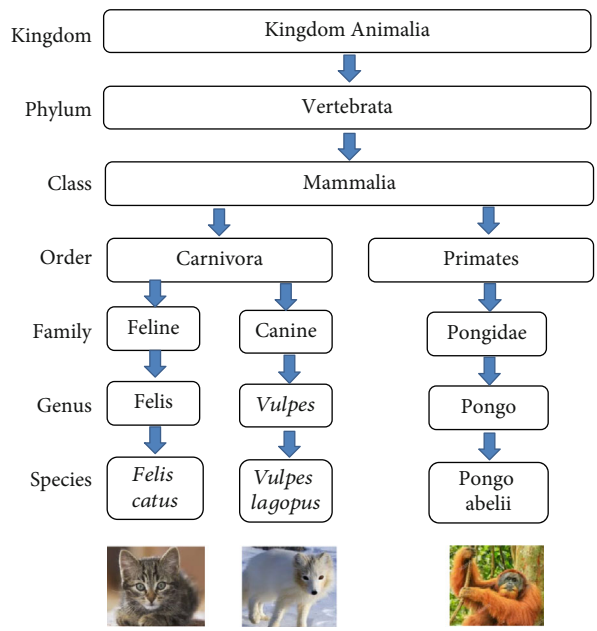


Figure 1: An example of a taxonomy of biology.

However, to the best of our knowledge, there is no related work taking consideration of both two limitations. In this paper, given a quantitative sequence database D (with an external utility table), a user-defined minimum utility threshold, and a series of taxonomies denoting the hierarchical relation, we are committed to finding all high-utility sequences consisting of items with the hierarchical relation (i.e., high-utility hierarchical sequential patterns). In fact, mining such patterns is more complicated than high-utility sequential pattern mining and sequential pattern mining with hierarchical relation. Firstly, compared with high-utility sequential pattern mining, the introduction of hierarchical relation leads to consumption of a large amount of memory and has long execution times due to the combinatorial explosion of the search space. Secondly, the methods of mining sequential pattern mining with hierarchical relation cannot be directly applied, for the download closure property (also known as the Apriori property) [28] is not held under the utility-based framework.

To address the above issues, we propose a new algorithm called MHUH (mining high-utility hierarchical sequential patterns) to mine high-utility hierarchical sequential patterns (to be defined later) by taking several strategies. The major contributions of this paper are as follows.

Firstly, we introduce the concepts of hierarchical relation into high-utility sequential pattern mining and formulate the problem of high-utility hierarchical sequential pattern mining (HUHSPM). Especially, important concepts and components of HUHSPM are defined.

Secondly, we propose a two-phase algorithm named MHUH (mining high-utility hierarchical sequential patterns), the first algorithm for high-utility hierarchical sequential pattern mining. So that the underlying informative knowledge of hierarchical relation between different items will not be missed and to improve efficiency of extracting

HUHPs, several strategies (i.e., FGS, PBS, and Reduction) and a novel upper bounder TSWU are proposed.

Thirdly, substantial experiments were conducted on both real and synthetic datasets to assess the performance of the two-phase algorithm MHUH in terms of runtime, number of patterns, and scalability. In particular, the experimental results demonstrate that MHUH can extract more interesting patterns with underlying informative knowledge efficiently in HUHSPM.

The rest of this paper is organized as follows. Related work is briefly reviewed in Section 2. We describe the related definitions and problem statement of HUHSPM in Section 3. The proposed algorithm is presented in Section 4, and an experimental evaluation assessing the performance of the proposed algorithm is shown in Section 5. Finally, a conclusion is drawn and future work is discussed in Section 6.

2. Related Work

In this section, related work is discussed. The section briefly reviews (1) the main approaches for sequential pattern mining, (2) the previous work of high-utility sequential pattern mining, and (3) state-of-the-art algorithms for sequential pattern mining with hierarchical relation.

2.1. Sequential Pattern Mining. Agrawal et al. [28] first presented a novel algorithm Apriori holding the download closure property for association rule mining. The proposed Apriori algorithm is based on a candidate generation approach that repeatedly scans the database to generate and count candidate sequential patterns and prunes those infrequent. They then defined the problem of sequential pattern mining over a large database of customer transactions and proposed the efficient algorithms AprioriSome and AprioriAll [4]. Srikant and Agrawal then proposed GSP, which is similar to AprioriAll in the execution process but greatly improves performance over AprioriAll. As Apriori's successor, adopting several technologies including time constraints, sliding time windows, and taxonomies, GSP uses a multiple-pass, candidate generation-and-test method to find sequential patterns [6]. Zaki [1] proposed the efficient SPADE which only needs three database scans. SPADE utilizes combinatorial properties to decompose the original problem into smaller subproblems, which can be independently solved in main memory using efficient lattice search techniques and using simple join operations. Later, SPAM was proposed by Ayres et al. [29], which applies to the situation that sequential patterns in the database are very long. In order to deal with the problems of large search spaces and the ineffectiveness in handling dense datasets, Yang et al. [30] proposed a novel algorithm LAPIN with a simple idea that the last position is the key to judging whether to extend the candidate sequential patterns or not. Then, they developed the LAPIN-SPAM algorithm by combining SPAM, which outperforms SPAM up to three times on all kinds of dataset in experiments. Notably, the property, used in SPAM, LAPIN, LAPIN-SPAM, and SPADE, that the support of superpatterns is always less than or equal to the support of its support patterns is different from the Apriori property used in GSP. Summarizing all algorithms mentioned above, they all belong to Apriori-based algorithms [2, 3].

It is known that database scans will be time-consuming when discovering sequential patterns. For this reason, a set of pattern growth sequential pattern mining algorithms that are able to avoid recursively scanning the input data were proposed. For example, Han et al. [31] proposed a novel, efficient algorithm FreeSpan that uses projected sequential databases to confine the search and growth of subsequences. The projected database can greatly reduce the size of a database. Pei et al. [7] designed a novel data structure called web access pattern tree, or WAP-tree in short, in their algorithm. WAP-tree stores highly compressed and critical information, and it makes mine access patterns from web logs efficiently. Then, Han et al. [32] proposed PrefixSpan with two kinds of database projections level-by-level projection and bilevel projection. PrefixSpan projects only their corresponding postfix subsequences into the projected database, so it runs considerably faster than both GSP and FreeSpan. Using a preorder linked, position-coded version of WAP-tree and providing a position code mechanism, the PLWAP algorithm was proposed by Ezeife et al. based on WAP-tree [33]. Recently, Sequence-Growth, the parallelized version of the PrefixSpan algorithm, was proposed by Liang and Wu [34], which adopts a lexicographical order to generate candidate sequences that avoid exhaustive search over the transaction databases.

There are some drawbacks to pattern growth sequential pattern mining algorithms. Obviously, it is time-consuming to build projected databases. Consequently, some algorithms with early pruning strategies were developed to improve efficiency. Chiu et al. [35] designed an efficient algorithm called DISC-all. Different with previous algorithms, DISC-all adopts the DISC strategy to prune the nonfrequent sequences according to the other sequences with the same length instead of the frequent sequences with shorter lengths. Recently, a more fast algorithm called CloFAST was proposed for mining closed sequential patterns using sparse and vertical id-lists. CloFAST combines a new data representation of the dataset, whose theoretical properties are studied in order to fast count the support of sequential patterns, with a novel one-step technique both to check sequence closure and to prune the search space [36]. It is more efficient than previous approaches. More details about the background of sequential pattern mining can be found in [2, 3].

2.2. High-Utility Sequential Pattern Mining. To address the problem that frequency does not fully reveal the importance in many situations, utility-oriented pattern mining frameworks, for example, high-utility itemset mining (HUIM), have been proposed and extensively studied [12, 37]. Although HUIM algorithms can extract interesting patterns in many real-life applications, they are not able to handle the sequence database where the timestamp is embedded in each item. Many high-utility sequential pattern mining algorithms have been proposed in the past decades [9, 13, 16, 18, 38], and high-utility sequential patterns can be extracted more efficiently with a series of novel data structures and

pruning strategies proposed. Ahmed et al. [13] first defined the problem of mining high-utility sequential patterns and proposed a novel framework for mining high-utility sequential patterns. They presented two new algorithms UL and US to find all high-utility sequential patterns. The UL algorithm, which is simpler and more straightforward, follows the candidate generation approach (based on breadth-first search), while the US algorithm follows the pattern growth approach (based on depth-first search). They can both be regarded as two-phase algorithms. In the first phase, they find a set of high-SWU sequences. In the second phase, they calculate the utility of sequences by scanning the sequence database to output high-SWU sequences, only those whose utility is no less than the threshold *minutil*.

The two-phase algorithms mentioned above have two important limitations, especially for low *minutil* values [8]. One limitation is that the set of high-SWU sequences discovered in the first phase needs a considerable amount of memory. The other one is that computing the utility of candidate sequences can be very time-consuming when scanning the sequence database. Instead of dividing the algorithm into two phases, Shie et al. [22] proposed a one-phase algorithm named UM-Span for high-utility sequential pattern mining. It improves efficiency by using a projected database-based approach to avoid additional scans of databases to check actual utilities of patterns. Similarly, a one-phase algorithm named PHUS was proposed by Lan and Hong [39], which adopted an effective upper bound model and an effective projection-based pruning strategy. Furthermore, the indexing strategy is also developed to quickly find the relevant sequences for prefixes in mining, and thus, unnecessary search time can be reduced.

Yin et al. then enriched the related definitions and concepts of high-utility sequential pattern mining. Two algorithms, USpan [9] and TUS [17], were proposed by Yin et al. for mining high-utility sequential patterns and top-k high-utility sequential patterns, respectively. In USpan, they introduced the lexicographic quantitative sequence tree to represent the search space and designed concatenation mechanisms for calculating the utility of a node and its children with two effective pruning strategies. The width pruning strategy avoids constructing unpromising patterns into the LP-Tree, while the depth pruning strategy stops USpan from going deeper by identifying the leaf nodes in the tree. Based on USpan, Alkan and Karagoz and Wang et al., respectively, proposed HuspExt [16] and HUS-Span [38] to increase efficiency of the mining process. Zhang et al. [18] proposed an efficient algorithm named FHUSpan (named HUS-UT in the paper), which adopts a novel data structure named Utility-Table to store the sequence database in the memory and the TRSU strategy to reduce search space. Recently, Gan et al. proposed two efficient algorithms named ProUM [40] and HUSP-ULL [41], respectively, to improve mining efficiency. The former utilizes the projection technique in generating utility array, while the latter adopts a lexicographic q-sequence- (LQS-) tree and a utility-linked-list structure to quickly discover HUSPs. More current development of HUSPM can be referred to in literature reviews [8, 14].

2.3. Sequential Pattern Mining with Hierarchical Relation. Sequential pattern mining with hierarchical relation can be traced back to article [6] where the hierarchies were incorporated into the mining process, and the GSP algorithm was proposed to extract sequential patterns according to different levels of hierarchy. There are two key strategies to improve efficiency in GSP. The first one is precomputing the ancestors of each item and dropping ancestors which are not in any of the candidates before making a pass over the data. The second strategy is to not count sequential patterns with an element containing both item and its ancestor. However, the depth of the hierarchy limits the efficiency of the algorithm because it increases the size of the sequence database. To represent the relationships among items in a more complete and natural manner, Chen and Huang [25] sketched the idea of fuzzy multilevel sequential patterns and presented the FMSM algorithm and the CROSS-FMSM algorithm based on GSP. Each item in hierarchies can have more than one parent with different degrees of confidence in their paper.

Plantevit et al. [24] incorporated the concept of hierarchy into a multidimensional database and proposed the two-phase algorithm HYPE extending their preceding M^2SP approach to extract multidimensional h-generalized sequential patterns. Firstly, the maximally specific items are extracted. Secondly, the multidimensional h-generalized sequences are mined in a further step. As the accessor of HYPE, they then proposed M^3SP [42] to extract multidimensional and multilevel sequential patterns based on M^2SP. The approaches are not incomplete; in other words, they do not mine all frequent sequences. Similarly applying fuzzy concepts to the hierarchy, Huang [43] later presented a divide-and-conquer strategy based on the pattern growth approaches to mine such fuzzy multilevel patterns. Recently, Egho then presented MMISP to extract heterogeneous multidimensional sequential patterns with hierarchical relation and applied it to analyze the trajectories of care for colorectal cancer. Beedkar et al. [23], who were inspired by MG-FSM, designed the first parallel algorithm named LASH for efficiently mining sequential patterns with hierarchical relation. MG-FSM first partitions the data and subsequently mines each partition independently and in parallel. Drawing lessons from the basic strategy of MG-FSM, Lash adopts a novel, hierarchy-aware variant of item-based partitioning, optimized partition construction techniques and an efficient special-purpose algorithm called pivot sequence miner (PSM) for mining each partition. As we know, the sequence database contains not only rich features (e.g., occurrence quantity, risk, and profit) but also multidimensional auxiliary information, which is partly associated with the concept of hierarchy. Recently, Gan et al. [44] proposed a novel framework named MDUS to extract multidimensional utility-oriented sequential useful patterns.

There are also several hierarchical frequent itemset mining algorithms, which are more or less similar to sequential pattern mining with hierarchical relations. For example, Kiran et al. [45] proposed a hierarchical clustering algorithm using closed frequent itemsets that use Wikipedia as an external knowledge to enhance the document representation. In Prajapati and Garg's research [46], the transactional dataset

is generated from a big sales dataset; then, the distributed multilevel frequent pattern mining algorithm (DMFPM) is implemented to generate level-crossing frequent itemset using the Hadoop Mapreduce framework. And then, the multilevel association rules are generated from frequent itemset.

3. Preliminaries and Problem Formulation

3.1. Definitions. Let I be a set of items. A nonempty subset $X \subseteq I$ is called an itemset, and the symbol $|X|$ denotes the size of X. A sequence $S: <X_1, X_2, \cdots, X_n>$ is an ordered list of itemsets, where $X_k \subseteq I$ $(1 \le k \le n)$. The length of S is $\sum_{k=1}^{n} |X_k|$, and the size of S is n. A sequence with the length of l is called an l-sequence. $T: <Z_1, Z_2, \cdots, Z_m>$ is the subsequence of S, if there exists m integers: $1 \le k_1 < k_2 < \cdots < k_m \le n$ so that $\forall 1 \le v \le m, Z_v \subseteq X_{k_v}$. For example, $<\{a\}\{b\,c\}>$ is the subsequence of $<\{a\,b\}\{a\,b\,c\}\{b\}>$.

A q-item (quantitative-item) is an ordered tuple (i, q), where $i \in I$ and q is a positive real number representing the quantity of i. A q-itemset with n q-items is denoted as $\{(i_1, q_1)(i_2, q_2) \cdots (i_n, q_n)\}$. A q-sequence, denoted as $<Y_1, Y_2, \cdots, Y_m>$, is an ordered list of m q-itemsets. A q-sequence database D (e.g., Figure 2(a)) consists of a collection of tuple $<\mathrm{ID}, Q>$, where ID is the identifier and Q is a q-sequence.

The hierarchical relation of different items is represented in the form of taxonomy which is a tree consisting of items in different abstraction levels. We assume that each item is only associated with one taxonomy. Figure 2(b) shows a simple example of taxonomies. In a taxonomy, if an item i is an ancestor of item j, we say that i is more general than j/j is more specific than i, denoted as $i \prec_I j / j \succ_I i$. We distinguish three different types of items: leaf items (most specific, no descendants), root items (most general, no ancestors), and intermediate items. The complete set consisting of descendants of item i is denoted as $\mathrm{down}(i)$. For example, in Figure 2(b), A is a root item, a_2 is an intermediate item, $a_{2,1}$ is a leaf item, and $\mathrm{down}(a_2) = \{a_{2,1}\, a_{2,2}\, a_{2,3}\}$. In this paper, we assume that different items belonging to the same itemset/q-itemset belong to different taxonomies.

Given two itemsets X and Y, we say that X is more specific than or equal to Y/Y is more general than or equal to X (denoted as $X \succcurlyeq_{\mathrm{IS}} Y / Y \preccurlyeq_{\mathrm{IS}} X$), *iff* $|X| = |Y|$ and $\forall i \in X, \exists j \in Y$ so that $i \succ_I j$ or $i = j$. For example, in Figure 2(b), $\{a_{1,1}\, b_1\, C\} \succcurlyeq_{\mathrm{IS}} \{a_1\, B\, C\}$; $\{A\,B\} \succcurlyeq_{\mathrm{IS}} \{A\,B\}$. Similarly, given two sequences with the size of m, S, and T, we say that S is more specific than or equal to T/T is more general than or equal to (denoted as $S \succcurlyeq_S T / T \preccurlyeq_S S$); if $\forall 1 \le k \le m$, we have $X \succcurlyeq_{\mathrm{IS}} Y$, where X is the kth itemset of S and Y is the kth itemset of T. In particular, if $S \succcurlyeq_S T$ and $S \ne T$, we say that S is more specific than T/T is more general than S, denoted as $S \succ_S T / T \prec_S S$. For example, in Figure 2(b), $<\{a_{1,1}\, b_1\, C\}\{A\,B\}> \succ_S <\{a_1\, B\, C\}\{A\,B\}>$.

3.2. Utility Calculation. Each item i is associated with an external utility (represented as $\mathrm{eu}(i)$) which is a positive real number representing the weight of i. For a nonleaf item i, it should meet the condition that $\mathrm{eu}(i) \ge \max\{\mathrm{eu}(j) \mid j \in \mathrm{down}(i)\}$. The external utility of each item $i \in I$ is recorded in an external utility table (e.g., Figure 2(c)).

The utility of a q-item (i, q) is defined as $q \times \mathrm{eu}(i)$. The utility of a q-itemset/q-sequence/q-sequence database is the sum of the utility of the q-items/q-itemset/q-sequence it contains. For example, in Figure 2, the utility of $a_{2,1}$ in the 1st itemset of Q_4 is 6 (1×6); the utility of the 1st itemset of Q_4 is 16 $(6 + 10)$; the utility of Q_4, represented as $u(Q_4)$, is 44 $(16 + 4 + 20 + 4)$; and the utility of D_E, represented as $u(D_E)$, is 228 $(74 + 39 + 71 + 44)$.

Given an itemset $X: \{i_1, i_2, \cdots, i_m\}$ and a q-itemset : $\{(j_1, q_1)(j_2, q_2) \cdots (j_n, q_n)\}(m \le n)$, we say that X occurs in Y, denoted as $X \subset_{\mathrm{IS}} Y$, iff there exist m distinct integers: $1 \le k_1, k_2, \cdots, k_m \le n$ so that $\forall 1 \le v \le m, i_v = j_{k_v}$ or $i_v \prec_I j_{k_v}$. The utility of X in Y is defined as $u(X, Y) = \sum_{v=1}^{m} q_{k_v} \times \mathrm{eu}(j_{k_v})$ if $X \subset_{\mathrm{IS}} Y$; otherwise, $u(X, Y) = 0$. For example, in Figure 2, let $Y_1 = \{(a_2, 2)(C, 2)\}$, $\{a_2\} \subset_{\mathrm{IS}} Y_1$; $\{A\,C\} \subset_{\mathrm{IS}} Y_1$; $\{A\,B\} \not\subset_{\mathrm{IS}} Y_1$; $u(\{A\}, Y_1) = 14 (2 \times 7)$; $u(\{A\,C\}, Y_1) = 32 (14 + 18)$; and $u(\{A\,B\}, Y_1) = 0$.

Given a sequence $S: <X_1, X_2, \cdots, X_m>$ and a q-sequence $Q: <Y_1, Y_2, \cdots, Y_n>$ where $m \le n$, we make the following definitions. We say that S occurs in Q (denoted as $S \subset_S Q$) at position p: $<k_1, k_2, \cdots, k_m>$ iff there exist m integers: $1 \le k_1 < k_2 < \cdots < k_m \le n$ so that $\forall 1 \le v \le m, X_v \subset_{\mathrm{IS}} Y_{k_v}$. The utility of S in Q at p, denoted as $u(S, p, Q)$, is defined as $u(S, p, Q) = \sum_{v=1}^{m} u(X_v, Y_{k_v})$. For example, in Figure 2, $S_1: <\{A\}\{A\,C\}>$ occurs in Q_1 at position $<1, 3>$; $u(S_1, <1, 3>, Q_1) = 8 + 32 = 40$.

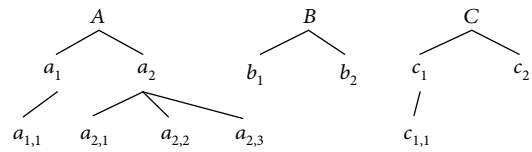
Obviously, S may occur in Q many times. The utility of S in Q, denoted as $u(S, Q)$, is defined as $u(S, Q) = \max\{u(S, p, Q) \mid p \in P(S, Q)\}$, where the symbol $P(S, Q)$ denotes the complete set containing all positions of S in Q. The utility of S in a q-sequence database D, denoted as $u(S)$, is defined as $u(S) = \sum_{Q \in D \wedge S \subset_S Q} u(S, Q)$. For example, in Figure 2, $S_2: <\{a_1\}\{C\}>$ occurs in Q_1 three times; $P(S_2, Q_1) = \{<1, 3>, <1, 4>, <1, 5>\}$; $u(S_2, <1, 3>, Q_1) = 8 + 18 = 26$; $u(S_2, Q_1) = \max\{26, 13, 16\} = 26$; $u(S_2) = 26 + 31 = 57$. More details about the methods of utility calculation can be found in [8].

Given a minimum utility ξ, we say that sequence W is high-utility if $u(W) \ge \xi$. In particular, W is the most specific pattern, denoted as s-sequence, iff $u(W) \ge \xi$ and $\forall T \succ_S W$, $u(T) < \xi$. Similarly, sequence S is the most general pattern, denoted as g-sequences, iff $u(S) \ge \xi$ and $\neg \exists T \prec_S S$. The s-sequences contain the underlying informative knowledge of hierarchical relations between different items, which cover less meaningless information compared with those sequences highly generalized. Therefore, we define these s-sequences as high-utility hierarchical sequential patterns (HUHSPs) to be extracted.

3.2.1. Problem Statement. Given a minimum utility ξ, a utility hierarchical sequence database including a quantitative sequential database D, a set of taxonomies, and an external utility table, the utility-driven mining problem of high-utility hierarchical sequential pattern mining (HUHSPM) consists of enumerating all HUHSPs whose overall utility

ID	q-sequence
Q_1	$<\{(a_{1,1}, 1)\}\ \{(a_2, 3)\}\ \{(a_2, 2)(C, 2)\}\ \{(c_{1,1}, 1)\}\ \{(c_2, 4)\}>$
Q_2	$<\{(a_{1,1}, 2)\}\ \{(a_2, 1)\}\ \{(b_2, 2)\}\ \{(b_1, 4)\}>$
Q_3	$<\{(a_{1,1}, 2)\}\ \{(c_{1,1}, 3)\}\ \{(A, 4)\}>$
Q_4	$<\{(a_{2,1}, 1)(c_{1,1}, 2)\}\ \{(a_{2,2}, 2)\}\{(c_{1,1}, 4)\}\ \{(a_{2,3}, 4)\}>$

(a) Quantitative-sequence database DE



(b) Taxonomies

Item	$a_{1,1}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	a_1	a_2	A	b_1	b_2	B	$c_{1,1}$	c_1	c_2	C
eu	8	6	2	1	8	7	10	3	2	3	5	5	2	9

(c) External utility table

Figure 2: An example of a hierarchical utility sequence database.

values in this database are no less than the prespecified minimum utility account ξ.

4. Proposed HUHSPM Algorithm: MHUH

In this section, we present the proposed algorithm MHUH for HUHSPM. We incorporate the hierarchical relation of items into high-utility sequential pattern mining, which makes MHUH able to find the underlying informative knowledge of hierarchical relation between different items ignored in high-utility sequential pattern mining. In other words, MHUH can extract more interesting patterns. The mining process of MHUH mainly includes two phases named Extension and Replacement. MHUH finds high-utility sequences by the existing algorithm FHUSpan (also named HUS-UT) which we proposed earlier based on the prefix-extension approach in the first phase. For a g-sequence S, we then generate all sequences that are more specific than S by progressive replacement and store s-sequences with a collection G in the second phase. The work we need to do in the two phases can be observed visually from the two names. The mining process with two phases ensures that the underlying informative knowledge of hierarchical relation between different items will not be missed. At the same time, it can increase efficiency when discovering HUHSPs.

Without the loss of generality, in this section, we formalize the theorems under the context of a minimum utility ξ and a utility hierarchical sequence database (includes a q-sequence database D, taxonomies, and external utility table).

4.1. Reduction: Remove Useless Items. Before mining the sequential patterns, MHUH adopts the Reduction strategy in the data preprocessing procedure, which removes useless items to reduce search space in advance. It mainly consists of two points, removing the unpromising items from the q-sequence database and removing the redundant items from the taxonomies.

An item is unpromising if any sequence containing this item is not high-utility. Here, we propose a novel upper bound TSWU (Taxonomy Sequence-Weighted Utility) based on SWU [13] to filter out the unpromising items.

Definition 1. Given an item i, we define $\mathrm{TSWU}(i)$ as $\sum_{Q\in D\wedge<\{r\}>\subset_S Q} u(Q)$, where r is the root item of the taxonomy containing i.

For example, in Figure 2, $\mathrm{TSWU}(a_{1,1}) = \mathrm{TSWU}(A) = 228$;$\mathrm{TSWU}(b_1) = \mathrm{TSWU}(b_2) = \mathrm{TSWU}(B) = 39$; $\mathrm{TSWU}(c_2) = \mathrm{TSWU}(C) = 189$.

Theorem 2. *Given a q-sequence Q, two sequences S_1 and S_2, where $S_1 \succ_S S_2$, $P(S_1, Q) \subseteq P(S_2, Q)$.*

Proof. Let $S_1 : <X_1, X_2, \cdots, X_m>, S_2 : <Z_1, Z_2, \cdots, Z_m>$. We have $\forall 1 \le v \le m, Z_v \preccurlyeq_{\mathrm{IS}} X_v$. For a q-sequence $Q : <Y_1, Y_2, \cdots, Y_n> (m \le n)$, $\forall p : <k_1, k_2, \cdots, k_m> \in P(S_1, Q)$, $Z_v \preccurlyeq_{\mathrm{IS}} X_v \subset_{\mathrm{IS}} Y_{j_v}$ $(1 \le v \le m)$, so $p \in P(S_2, Q)$. Further, $P(S_1, Q) \subseteq P(S_2, Q)$.

Theorem 3. *For any sequence S that contains item i, $u(S) < \xi$ if $TSWU(i) < \xi$.*

Proof. From Theorem 3, we know that $P(<\{i\}>, Q) \subseteq P(<\{r\}>, Q)$, so $\{Q \mid Q \in D \wedge <\{i\}> \subset_S Q\} \subseteq \{Q \mid Q \in D \wedge <\{r\}> \subset_S Q\}$. If $\mathrm{TSWU}(i) < \xi$, $u(S) = \sum_{Q\in D\wedge S\subset_S Q} u(S, Q) \le \sum_{Q\in D\wedge S\subset_S Q} u(Q) \le \sum_{Q\in D\wedge<\{i\}>\subset_S Q} u(Q) \le \sum_{Q\in D\wedge<\{r\}>\subset_S Q} u(Q) = \mathrm{TSWU}(i) < \xi$.

For a given ξ, we can remove items satisfying $\mathrm{TSWU}(i) < \xi$ safely according to the above theorem. For example, in Figure 2, when $\xi = 40$, b_1 and b_2 can be safely removed from Figure 2(a).

We say that an item is redundant if it (1) appears in taxonomy but does not appear in q-sequence database and (2) has at most one child in taxonomy. For example, in Figure 2, a_1 and c_1 are redundant items. In terms of utility, removing these items has no effect on correctness, which will be proved in Subsection 4.3. Thus, we can safely remove these items.

4.2. Extension (Phase I): Find g-Sequences. In the first phase named Extension, we use the existing algorithm FHUSpan [18] which we proposed earlier to efficiently mine the general high-utility sequences (g-sequences). The main tasks of this phase are improving efficiency greatly of MUHU and extract g-sequences preparing for the next phase.

In fact, no s-sequences will be missed based on the FGS (From General to Special) strategy. To prove the correctness of this conclusion, we need to prove two points: (1) there does not exist s-sequence S that cannot be discovered by the FGS strategy and (2) the correctness of the algorithm that finds s-sequence is based on a given g-sequence. Here, we prove the correctness of (1), and the proof about (2) is illustrated in the next subsection.

Theorem 4. *Given two sequences S_1 and S_2 where $S_1 \succ_S S_2$, $u(S_1) \leq u(S_2)$.*

Proof. We first prove that $u(S_1, Q) \leq u(S_2, Q)$. From Section 4.2 ($\text{eu}(i) \geq \max\{\text{eu}(j) \mid j \in \text{down}(i)\}$) and Theorem 3, $u(S_2, Q) = \max\{u(S_2, p, Q) \mid p \in P(S_2, Q)\} \geq \max\{u(S_1, p, Q) \mid p \in P(S_1, Q)\} = u(S_1, Q)$. Then, we prove $u(S_1) \leq u(S_2)$. We have $D_1 \subseteq D_2$ if $S_1 \succ_S S_2$, where $D_k(k = 1, 2) = \{Q \mid Q \in D \wedge S_k \subset_S Q\}$. $u(S_2) = \sum_{Q \in D_2 \setminus D_1} u(S_2, Q) + \sum_{Q \in D_1} u(S_2, Q) \geq \sum_{Q \in D_1} u(S_2, Q) \geq \sum_{Q \in D_1} u(S_1, Q) = u(S_1)$.

Corollary 5. *Given a g-sequence S, $\forall T \succ_S S, u(T) \leq u(S)$.*

Theorem 4 and Corollary 5 reveal the correctness of (1). We assume that S is a s-sequence that cannot be discovered by the FGS strategy. In fact, we can always find (replace item with the item's ancestor) the sequence S_g where $S_g \preccurlyeq_S S$ and $\neg\exists T \prec_S S$. Because S_g is not g-sequence, $u(S_g) < \xi$ or $\exists T \prec_S S$. So, $u(S_g) < \xi$. We then draw a contraction that $S_g \preccurlyeq_S S \wedge u(S_g) < u(S)$. Therefore, the assumption does not hold, which ensures the correctness of (1).

Theorem 6. *All items contained in g-sequence are root items.*

Proof. Given a g-sequence S, we assume that the ith item of S is not the root item. Then, we can find a sequence T where $T \prec_S S$. We then draw a contraction that S is not g-sequence. Therefore, the theorem holds.

We then introduce how to find g-sequences. Theorem 6 shows that we merely need to consider the root items in the process of finding g-sequences. Thus, we can transform the q-sequences into another form so that we can ignore the hierarchical relation in this phase. We illustrate this transformation through an example. Consider Q_3 in Figure 2, we transform it into $<\{A[16]\}\{C[15]\}\{A[40]\}>$, where the value in the bracket is utility ($\text{eu} \times \text{iu}$). Obviously, with this transformation, mining g-sequence is equivalent to mining high-utility sequences. So, we use the existing high-utility sequential pattern mining algorithm FHUSpan [18], which we proposed earlier to find g-sequences.

Here, we briefly introduce the mining process of FHUSpan, which finds high-utility sequences based on the prefix-extension approach. It first finds all appropriate items (only the sequence starting with these items may be high-utility). Then, for each appropriate item, it constructs a sequence containing only this item and extends the sequence recursively until all sequences starting with the item are checked. In particular, two extension approaches are used, S-Extension (appending an itemset containing only one item to the end of the current sequence) and I-Extension (appending an item to the last itemset of the current sequence). It is based on the algorithm HUS-Span which uses two pruning strategies, PEU (Prefix Extension Utility) strategy and RSU (Reduced Sequence Utility) strategy to reduce the search space. The novel data structure named Utility-Table and the pruning are used to terminate extension in FHUSpan so that it can efficiently discover high-utility sequences.

4.3. Replacement (Phase II): Find s-Sequence. In the second phase named Replacement, we mine the special high-utility sequences with the hierarchical relation (s-sequences) from g-sequences by the PBS strategy. The main task of this phase is to extract s-sequences efficiently.

For a g-sequence S, we then generate all sequences that are more specific than S by progressive replacement and store s-sequences with a collection G. In particular, for each replacement, we replace the kth item of S with a child item. For example, in Figure 2, we replace the first item of $S_3 : <\{A\}\{A\}\{C\}>$ with the child items of A, and one specific sequence is $S_4 : <\{a_2\}\{A\}\{C\}>$.

Algorithm 1 shows the progressive replacement starting from the kth item of a sequence, which is based on DFS. Firstly, it checks if the current sequence S has been visited to avoid repeated utility calculation (line 1). If $u(S) < \xi$, we have $\forall S' \succ_S S, u(S') < \xi$ according to Theorem 4, so we terminate search (lines 2-3). Otherwise, it adds S into G and removes the sequences that are more general than S from G (line 5). Then, it generates the more specific sequences based on S, which follows the order from top to bottom (line 9), left to right (lines 10-12). In detail, it first finds $R(S, k)$ which is the set containing all child items for replacement. For each $r \in R(S, k)$, it replaces the ith item of S with r to generate S'. It then checks the sequences that are more specific than S' (line 9, from top to bottom). After that, it checks the sequences that are more specific than S' from left to right (lines 10-12), where l is the length of S'.

We also use a strategy, PBS (Pruning Before Search), to reduce search space before Algorithm 1. The main idea behind this strategy is considering only the items in the current index. In other words, we generate and check the more specific sequences in one direction (from top to bottom) to reduce the size of taxonomies.

We illustrate this strategy through an example under the context of Figure 2. Let $\xi = 45$, the sequence $S_3 : <\{A\}\{A\}\{C\}>$ is a g-sequence ($u(S_3) = 77 > \xi$). We construct copies of taxonomy, denoted as T_1, T_2, and T_3, for the kth ($k = 1, 2, 3$) item of S_3. Then, we reduce the size of the three taxonomies. For T_1, we have $R(S_3, 1) = \{a_{1,1}, a_2\}$ (a_1 is a redundant item and was removed). Then, we generate $S_5 : <\{a_{1,1}\}\{A\}\{C\}>$ by replacing the first A with $a_{1,1}$, and $u(S_5) = 47 > \xi$. So, we retain $a_{1,1}$. Because $R(S_5, 1) = \varnothing$, we then consider a_2 and generate $S_6 : <\{a_2\}\{A\}\{C\}>$. We also retain a_2, for $u(S_6) = 73 > \xi$. Then, we continue to check the child items of a_2. Such a procedure will continue until all items belonging to $\text{down}(A)$ have been checked. Finally, we

```
Search for specific.
Input: S: the sequence, k: start index, visited, G
if visitedisfalse then
    if u(S) < ξ then
        return
    end if
        G ⟵ Filter(G ∪ S)
end if
for all r ∈ R(S,k) do
        S' ⟵ Replace(S,k,r)
        SearchForSpecific(S ,k,false,G)
    for v = k +1 ⟶ l do
            SearchForSpecific(S ,v,true,G)
    end for
end for
```

Algorithm 1

remove $a_{2,1}, a_{2,2}, a_{2,3}$ from T_1, for $u(<\{a_{2,1}\}\{A\}\{C\}>) = 30 < \xi$, $u(<\{a_{2,2}\}\{A\}\{C\}>) = u(<\{a_{2,3}\}\{A\}\{C\}>) = 0 < \xi$. We then continue the above procedure for T_2 and T_3, and the processed taxonomies are shown in the right of Figure 3. In addition, note that in Algorithm 1, $R(S, k)$ is obtained from the processed taxonomies instead of the original taxonomies.

In the above example, the max count of sequences that are more specific than S reduces from 107 $(6 \times 6 \times 3 - 1)$ to 11 $(3 \times 2 \times 2 - 1)$. In fact, for a l-sequence S, this count reduces from $\prod_{k=1}^{l}(d_k^1 + 1) - 1$ to $\prod_{k=1}^{l}(d_k^2 + 1) - 1$, where d_k^1 and d_k^2 are the sizes of down(i_k) in the original and processed taxonomies, respectively, and i_k is the kth item of S.

In the rest of this subsection, we prove the conclusion left before. We first prove that removing redundant items has no effect on correctness.

Proof. For a sequence S, we assume that the kth item of S, i_k, is a redundant item. Firstly, if i_k is a leaf item, we can safely remove it, because for each sequence W that contains i_k, we have $u(W) = 0$. Secondly, if i_k has one child, we generate sequence W by replacing i_k with its child. Then, we have $u(W) = u(S)$ according to the related utility definition (the utility of X in Y). Therefore, removing redundant items does not change the utility of related sequences, which means that it has no effect on the correctness.

Then, we prove the conclusion the correctness of the algorithm which finds s-sequence based on a given g-sequence.

Proof. Firstly, the PBS strategy does not ignore the underlying s-sequences. Suppose we cannot find a s-sequence S from the taxonomies processed by PBS strategy, then we have $\exists S_g \prec_S S, u(S_g) < u(S)$, which violates Theorem 4. So, the assumption does not hold. Secondly, Algorithm 1 does not ignore any s-sequences. Algorithm 1 is based on the DFS framework, which ensures the completeness of the algorithm. Besides, Algorithm 1 terminates search in advance based on Theorem 4, so it does not ignore any s-sequences. In summary, the conclusion holds.

5. Experiments

We performed experiments to evaluate the proposed MHUH algorithm which was implemented in Java. All experiments were carried out on a computer with Intel Core i7 CPU of 3.2 GHz, 8 GB memory, and Windows 10.

5.1. Datasets. Five datasets, including three real datasets and two synthetic datasets, were used in the experiments. DS1 is the conversion of Bible where each word is an item. DS2 is the conversion of the classic novel called Leviathan. DS3 is a click-stream dataset called BMSWebView2. The three datasets can be obtained from the SPMF website [47]. DS4 and DS5 are two synthetic datasets. The characteristic of them is summarized in Table 1. The values of parameters in Table 1 are as follows: $\#S$ is the number of sequences, $\#I$ is the number of distinct items, M is the max length of sequences, and A is the average length of sequences.

Note that these datasets do not contain taxonomies. So, for each dataset, we generated taxonomies based on the items it contains. The max depth and degree of these taxonomies are 3, which indicates that the max number of leaf items contained in taxonomy is 27. The datasets and source code will be released at the author's Github after the acceptance for publication.

5.2. Performance Evaluation. We evaluated the performance of the proposed algorithm on different datasets when varying ξ. For the sake of simplicity, here, we calculate ξ as $\delta \times u(D)$, where δ is a decimal between 0 and 1, and $u(D)$ is the utility of the q-sequence database (see the concepts in Subsection 3.2). In addition, we also tested the effect of the PBS strategy, and the modified MHUH algorithm which does not take the PBS strategy is denoted as MHUH_base.

The execution times of MHUH and MHUH_base on DS1 to DS3 are shown in Figure 4. When δ/ξ increases, both of the two algorithms take less execution time since the search space reduces. The results prove that the PBS strategy effectively decreases the execution time, for it greatly reduces the search space on these datasets. Besides, the results also show that the MHUH algorithms can efficiently extract s-sequences under a low ξ.

Figure 5 shows the distribution of discovered patterns by MHUH on DS1 to DS3. It shows that the number of patterns per length increases with the decrease of ξ. In particular, it is interesting that some longer patterns may disappear as ξ increases, which indicates that the shorter patterns may have higher utility.

5.3. Utility Comparison with High-Utility Sequential Pattern Mining. We conducted this experiment to evaluate the utility difference between the patterns discovered by MHUH and that discovered by the existing algorithm FHUSpan [18] which we proposed earlier.

Figure 6 shows the sum utility of top # (depends on utility) patterns discovered by FHUSpan and MHUH from three datasets. The X-axis refers to the value of #, and the Y-axis

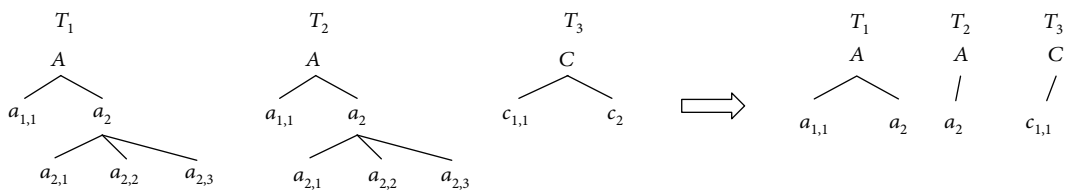


Figure 3: Reduce the size of T_1, T_2, and T_3 through the PBS strategy.

Table 1: Characteristic of datasets.

Dataset	#S	#I	M	A	Type
DS1	36,369	13,905	100	21.64	Real (text)
DS2	5,834	9,162	100	33.81	Real (text)
DS3	77,512	6,120	161	4.62	Real (click stream)
DS4	10,000	4,000	40	20.54	Synthetic
DS5	60,000	5,000	20	10.50	Synthetic

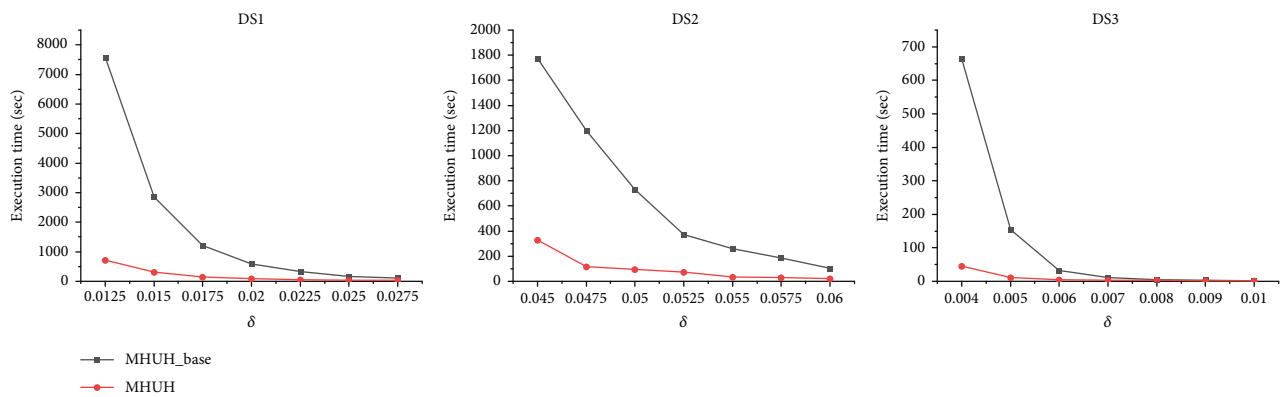


Figure 4: Execution time on three datasets when varying δ.

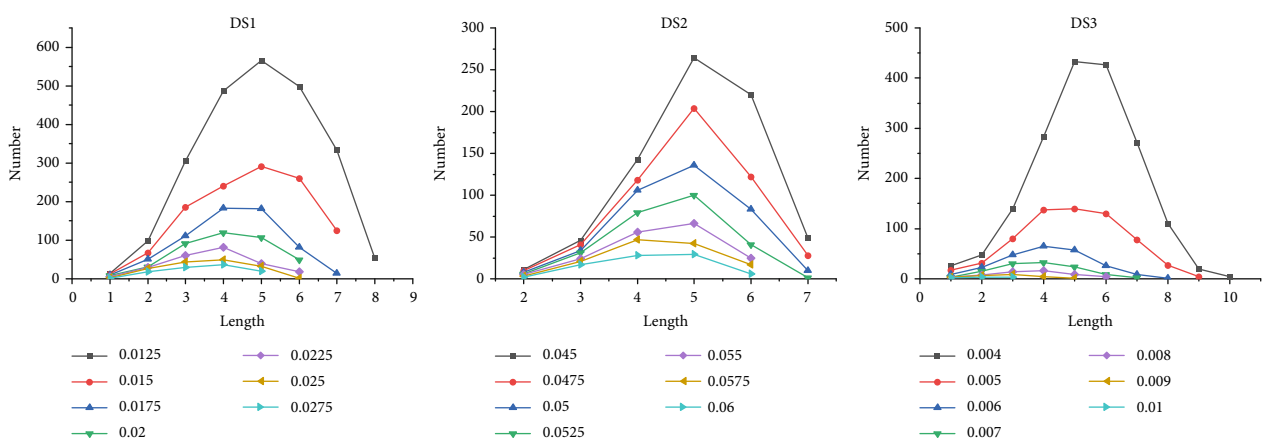


Figure 5: Distribution of discovered patterns on three datasets when varying δ.

represents the sum utility of top # patterns. For example, on DS1, the sum utility of top 1000 patterns extracted by MHUH is higher than the sum utility of that discovered by FHUSpan. Figure 7 shows that the average utility per length of top # patterns on DS1 to DS3 (# is set to 1000, 700, and 600, respectively). The X-axis refers to the length of patterns, and the Y-axis denotes the average utility of patterns with the same length. For example, on DS1, in terms of the top 1000 patterns, the average utility of patterns with length of 8 discovered by MHUH is higher than the average utility of that discovered by FHUSpan. From these two figures, we know that MHUH can discover higher utility patterns compared

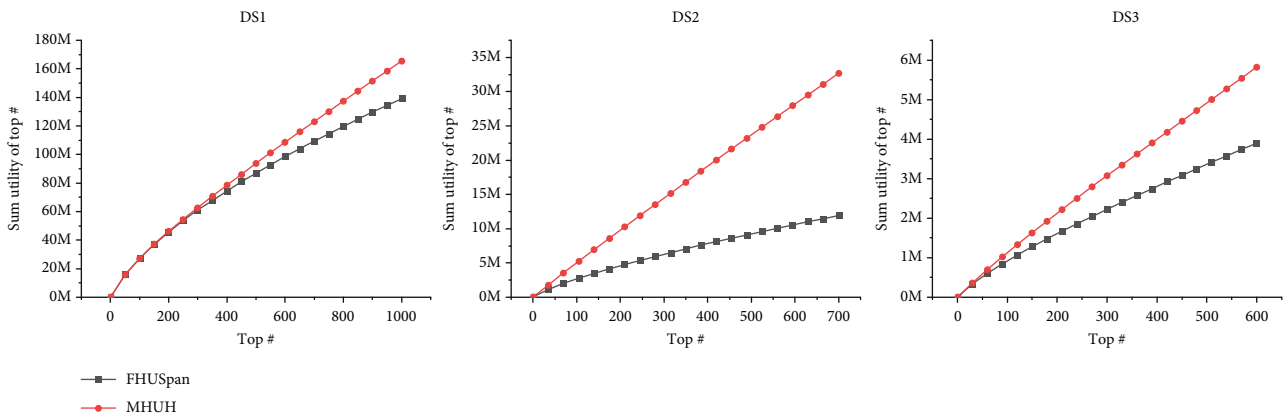


Figure 6: Sum utility of top # patterns discovered from three datasets.

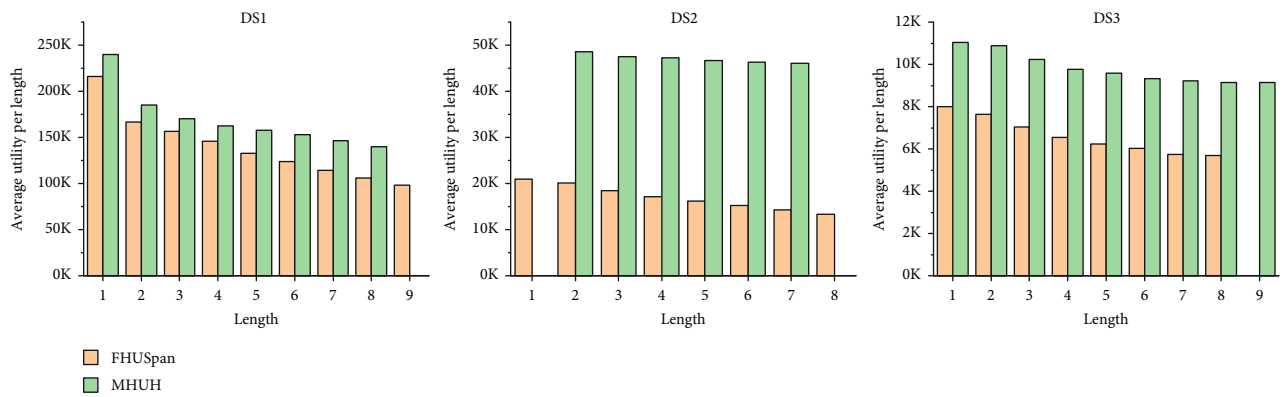


Figure 7: Average utility per length of top # patterns.

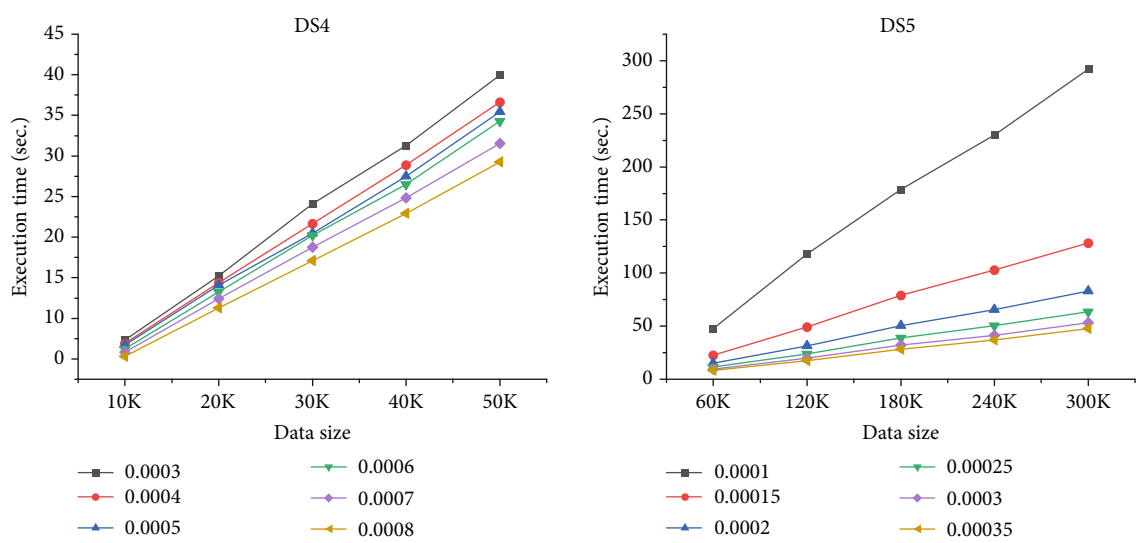


Figure 8: Scalability test on two datasets.

with FHUSpan, indicating that more informative knowledge can be found by MHUH.

5.4. Scalability. We conducted experiments to evaluate MHUH's performance on large-scale datasets. For each dataset, we increased its data size through duplication and performed the MHUH algorithm with different δ. Figure 8 shows the experimental results. We know from the figure that the MHUH algorithm has well scalability on the two datasets, for the execution time is almost linear with the data

size. For example, the execution time of MHUH ($\delta = 0.0003$) on DS4 almost linearly increases when the data size (the number of q-sequences it contains) changes from 10K to 50K. It also shows that MHUH can efficiently identify the desired patterns from the large-scale dataset with a low ξ. For example, in terms of DS5, MHUH costs 300 s when the data size is 300K and $\delta = 0.0001$.

6. Conclusion and Future Work

In this paper, we incorporate the hierarchical relation of items into high-utility sequential pattern mining and propose a two-phase algorithm MHUH, the first algorithm for high-utility hierarchical sequential pattern mining (HUHSPM). In the first phase named Extension, we use the existing algorithm FHUSpan which we proposed earlier to efficiently mine the general high-utility sequences (g-sequences); in the second phase named Replacement, we mine the special high-utility sequences with the hierarchical relation (s-sequences) from g-sequences. The proposed MHUH algorithm takes several novel strategies (e.g., Reduction, FGS, and PBS) and a new upper bound TSWU, so it will be able to greatly reduce the search space and discover the desired pattern HUHSPs efficiently. A conclusion can be drawn from the experiment that MHUH extracts more interesting patterns with underlying informative knowledge efficiently in HUHSPM.

In the future, we will generalize the proposed algorithm based on the more complete concepts. Besides, several extensions of the proposed MHUH algorithm can be considered such as improving the efficiency of the MHUH algorithm based on better pruning strategies, efficient data structures [40, 41], and the multithreading technology [2].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors have declared that no conflict of interest exists.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province, China (Grant No. 2020A15150109 70) and Shenzhen Research Council (Grant No. GJHZ2018 0928155209705).

References

[1] M. J. Zaki, "Spade: an efficient algorithm for mining frequent sequences," *Machine Learning*, vol. 42, no. 1/2, pp. 31–60, 2001.

[2] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, and P. S. Yu, "A survey of parallel sequential pattern mining," *ACM Transactions on Knowledge Discovery from Data*, vol. 13, no. 3, pp. 1–34, 2019.

[3] P. Fournier-Viger, J. C.-W. Lin, R. U. Kiran, Y. S. Koh, and R. Thomas, "A survey of sequential pattern mining," *Data Science and Pattern Recognition*, vol. 1, no. 1, pp. 54–77, 2017.

[4] R. Agrawal and R. Srikant, "Mining sequential patterns," in *ICDE '95: Proceedings of the Eleventh International Conference on Data Engineering*, vol. 95, pp. 3–14, Taipei, Taiwan, China, March 1995.

[5] H. Lu, Y. Li, S. Mu, D. Wang, H. Kim, and S. Serikawa, "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2315–2322, 2018.

[6] R. Srikant and R. Agrawal, "Mining sequential patterns: generalizations and performance improvements," in *International Conference on Extending Database Technology*, pp. 1–17, Springer, Berlin, Heidelberg, 1996.

[7] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 396–407, Springer, Berlin, Heidelberg, 2000.

[8] T. Truong-Chi and P. Fournier-Viger, "A survey of high utility sequential pattern mining," in *High-Utility Pattern Mining: Theory, Algorithms and Applications*, pp. 97–129, Springer, 2019.

[9] J. Yin, Z. Zheng, and L. Cao, "Uspan: an efficient algorithm for mining high utility sequential patterns," in *18th ACM SIGKDD International conference on Knowledge discovery and data mining*, pp. 660–668, Beijing, China, August 2012, ACM.

[10] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain intelligence: go beyond artificial intelligence," *Mobile Networks and Applications*, vol. 23, no. 2, pp. 368–375, 2018.

[11] W. Gan, J. Chun-Wei, H.-C. Chao, S.-L. Wang, and S. Y. Philip, "Privacy preserving utility mining: a survey," in *2018 IEEE International Conference on Big Data*, pp. 2617–2626, Seattle, WA, USA, Nov 2018, IEEE.

[12] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, T.-P. Hong, and H. Fujita, "A survey of incremental high-utility itemset mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 2, article e1242, 2018.

[13] C. F. Ahmed, S. K. Tanbeer, and B.-S. Jeong, "A novel approach for mining high-utility sequential patterns in sequence databases," *ETRI Journal*, vol. 32, no. 5, pp. 676–686, 2010.

[14] W. Gan, J. C.-W. Lin, P. Fournier-Viger, H.-C. Chao, V. S. Tseng, and P. S. Yu, "A survey of utility-oriented pattern mining," 2018, http://arxiv.org/abs/1805.10511.

[15] T. Wang, X. Ji, A. Song et al., "Output bounded and rbfnn-based position tracking and adaptive force control for security tele-surgery," *ACM Transactions on Multimedia Computing Communications and Applications*, ACM, 2020.

[16] O. K. Alkan and P. Karagoz, "Crom and huspext: Improving efficiency of high utility sequential pattern extraction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 10, pp. 2645–2657, 2015.

[17] J. Yin, Z. Zheng, L. Cao, Y. Song, and W. Wei, "Efficiently mining top-k high utility sequential patterns," in *2013 IEEE 13th International conference on data mining*, pp. 1259–1264, Dallas, TX, USA,, 2013, IEEE.

[18] C. Zhang, Z. Yiwen, J. Nie, and D. Zilin, "Two efficient algorithms for mining high utility sequential patterns," in *17th*

IEEE International Symposium on Parallel and Distributed Processing with Applications, Xiamen, China, 2019IEEE.

[19] H. Lu, D. Wang, Y. Li et al., "Conet: a cognitive ocean network," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 90–96, 2019.

[20] X. Yang, H. Wu, Y. Li et al., "Dynamics and isotropic control of parallel mechanisms for vibration isolation," in *IEEE/ASME Transactions on Mechatronics*, IEEE, 2020.

[21] C. F. Ahmed, S. K. Tanbeer, and B.-S. Jeong, "Mining high utility web access sequences in dynamic web log data," in *2010 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pp. 76–81, London, UK, June 2010, IEEE.

[22] B.-E. Shie, J.-H. Cheng, K.-T. Chuang, and V. S. Tseng, "A one-phase method for mining high utility mobile sequential patterns in mobile commerce environments," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 616–626, Berlin, Heidelberg, 2012.

[23] K. Beedkar and R. Gemulla, "Lash: large-scale sequence mining with hierarchies," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 491–503, Melbourne, Australia, May 2015, ACM.

[24] M. Plantevit, A. Laurent, and M. Teisseire, "Hype: mining hierarchical sequential patterns," in *Proceedings of the 9th ACM International workshop on Data warehousing and OLAP*, pp. 19–26, Arlington, Virginia, USA, November 2006, ACM.

[25] Y.-L. Chen and T. C.-K. Huang, "A novel knowledge discovering model for mining fuzzy multilevel sequential patterns in sequence databases," *Data & Knowledge Engineering*, vol. 66, no. 3, pp. 349–367, 2008.

[26] R. Srikant and R. Agrawal, "Mining generalized association rules," *Future Generation Computer Systems*, vol. 13, no. 2-3, pp. 161–180, 1997.

[27] L. V. Q. Anh and M. Gertz, "Mining spatio-temporal patterns in the presence of concept hierarchies," in *2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 765–772, Brussels, Belgium, Dec. 2012, IEEE.

[28] R. Agrawal R. Srikant et al., "Fast algorithms for mining association rules," in *Proceedings of the 20th VLDB Conference*, vol. 1215, pp. 487–499, Santiago, 1994.

[29] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proceedings of the eighth ACM SIGKDD International conference on Knowledge discovery and data mining*, pp. 429–435, Edmonton Alberta, Canada, July 2002, ACM.

[30] Z. Yang, Y. Wang, and M. Kitsuregawa, "Lapin: effective sequential pattern mining algorithms by last position induction for dense databases," in *Advances in Databases: Concepts, Systems and Applications*, pp. 1020–1023, Springer, 2007.

[31] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu, "Freespan: frequent pattern-projected sequential pattern mining," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 355–359, Boston Massachusetts USA, August 2000.

[32] J. Han, J. Pei, B. Mortazavi-Asl et al., "Prefixspan: mining sequential patterns efficiently by prefix-projected pattern growth," in *Proceedings of the 17th international conference on data engineering*, pp. 215–224, Heidelberg, Germany, Germany, April 2001.

[33] C. I. Ezeife, Y. Lu, and Y. Liu, "Plwap sequential mining: open source code," in *Proceedings of the 1st International workshop on open source data mining: frequent pattern mining implementations*, pp. 26–35, Chicago Illinois USA, August 2005.

[34] Y.-h. Liang and W. Shiow-yang, "Sequence-growth: a scalable and effective frequent itemset mining algorithm for big data based on mapreduce framework," in *2015 IEEE International Congress on Big Data*, pp. 393–400, New York, NY, USA, Jun 2015.

[35] D.-Y. Chiu, Y.-H. Wu, and A. L. P. Chen, "An efficient algorithm for mining frequent sequences by a new strategy without support counting," in *Proceedings of the 20th International Conference on Data Engineering*, pp. 375–386, Boston, MA, USA, USA, April 2004.

[36] F. Fumarola, P. F. Lanotte, M. Ceci, and D. Malerba, "Clofast: closed sequential pattern mining using sparse and vertical id-lists," *Knowledge and Information Systems*, vol. 48, no. 2, pp. 429–463, 2016.

[37] C. F. Ahmed, S. K. Tanbeer, Byeong-Soo Jeong, and Young-Koo Lee, "Efficient tree structures for high utility pattern mining in incremental databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 12, pp. 1708–1721, 2009.

[38] J.-Z. Wang, J.-L. Huang, and Y.-C. Chen, "On efficiently mining high utility sequential patterns," *Knowledge and Information Systems*, vol. 49, no. 2, pp. 597–627, 2016.

[39] G.-C. Lan, T.-P. Hong, V. S. Tseng, and S.-L. Wang, "Applying the maximum utility measure in high utility sequential pattern mining," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5071–5081, 2014.

[40] W. Gan, J. C.-W. Lin, J. Zhang, H.-C. Chao, H. Fujita, and P. S. Yu, "Proum: projection-based utility mining on sequence data," *Information Sciences*, vol. 513, pp. 222–240, 2020.

[41] W. Gan, J. C.-W. Lin, J. Zhang, P. Fournier-Viger, H.-C. Chao, and P. S. Yu, "Fast utility mining on sequence data," *IEEE Transactions on Cybernetics*, pp. 1–14, 2020.

[42] M. Plantevit, A. Laurent, D. Laurent, M. Teisseire, and Y. W. E. I. Choong, "Mining multidimensional and multilevel sequential patterns," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 1, pp. 1–37, 2010.

[43] T. C.-K. Huang, "Developing an efficient knowledge discovering model for mining fuzzy multi-level sequential patterns in sequence databases," *Fuzzy Sets and Systems*, vol. 160, no. 23, pp. 3359–3381, 2009.

[44] W. Gan, J. C.-W. Lin, J. Zhang et al., "Utility mining across multi-dimensional sequences," 2019, http://arxiv.org/abs/1902.09582.

[45] G. V. R. Kiran, R. Shankar, and V. Pudi, "Frequent itemset based hierarchical document clustering using Wikipedia as external knowledge," in *International Conference on Knowledge-based and Intelligent Information and Engineering Systems*, pp. 11–20, Cardiff, Wales, UK, September 2010.

[46] D. J. Prajapati and S. Garg, "Map reduce based multilevel association rule mining from concept hierarchical sales data," in *International Conference on Advances in Computing and Data Sciences*, pp. 624–636, Springer, Singapore, July 2017.

[47] P. Fournier-Viger, "SPMF: an open-source data mining library," June 2019, http://www.philippe-fournier-viger.com/spmf/.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8893494, 8 pages
https://doi.org/10.1155/2020/8893494

Research Article

Light Deep Model for Pulmonary Nodule Detection from CT Scan Images for Mobile Devices

Mehedi Masud,[1] Ghulam Muhammad,[2] M. Shamim Hossain,[3] Hesham Alhumyani,[1] Sultan S. Alshamrani,[1] Omar Cheikhrouhou,[1] and Saleh Ibrahim[4,5]

[1]*College of Computers and Information Technology, Taif University, Taif 21974, Saudi Arabia*
[2]*Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia*
[3]*Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia*
[4]*Electrical Engineering Department, Taif University, Saudi Arabia*
[5]*Computer Engineering Department, Cairo University, Egypt*

Correspondence should be addressed to Ghulam Muhammad; ghulam@ksu.edu.sa

Received 27 May 2020; Revised 11 June 2020; Accepted 13 June 2020; Published 3 July 2020

Academic Editor: Yin Zhang

The emergence of cognitive computing and big data analytics revolutionize the healthcare domain, more specifically in detecting cancer. Lung cancer is one of the major reasons for death worldwide. The pulmonary nodules in the lung can be cancerous after development. Early detection of the pulmonary nodules can lead to early treatment and a significant reduction of death. In this paper, we proposed an end-to-end convolutional neural network- (CNN-) based automatic pulmonary nodule detection and classification system. The proposed CNN architecture has only four convolutional layers and is, therefore, light in nature. Each convolutional layer consists of two consecutive convolutional blocks, a connector convolutional block, nonlinear activation functions after each block, and a pooling block. The experiments are carried out using the Lung Image Database Consortium (LIDC) database. From the LIDC database, 1279 sample images are selected of which 569 are noncancerous, 278 are benign, and the rest are malignant. The proposed system achieved 97.9% accuracy. Compared to other famous CNN architecture, the proposed architecture has much lesser flops and parameters and is thereby suitable for real-time medical image analysis.

1. Introduction

Due to the advancement of sophisticated machine learning algorithms, mobile computing, wireless communications [1, 2], and finally cognitive computing [3, 4], the healthcare industry is booming in recent years. The traditional healthcare industry is now gradually shifting towards the smart healthcare industry [5]. The smart healthcare enables patients to have their health problems diagnosed sitting at their homes, to get the prescription and advice online, and thereby to save time for communication and getting an appointment [6]. One of the major driving forces behind the rise of the smart healthcare industry is the invention of deep learning algorithms in machine learning domain [7]. Deep learning has brought about a paradigm shift to machine learning. For the last ten years, it was used in numerous applications for signal and image processing, including medical signals or images [8–10].

Lung cancer has become one of the leading causes of death worldwide. In a recent statistic of 2019 from the American Cancer Society, it was found that more than 142K people died of lung and bronchus cancer and more than 228K people were diagnosed for lung and bronchus cancer [11]. The number of fatal cancer deaths can be greatly reduced by early diagnosis.

The early detection of cancer can be detected in two ways: manually by radiologists or automatically by a computer-aided diagnosis (CAD) system. The CAD system is not a standalone system; it can only assist the radiologists or the

doctors to take a correct decision. The final decision depends on the radiologists or the doctors [12]. A radiologist needs careful observation of the density of pulmonary nodules, because at an early stage, this density may resemble the densities of other lung parts [13]. A CAD system tries to make a boundary of a pulmonary nodule by detecting some distinguishing features in the nodule. These features are either manual features or deep-learned features. Manual features include information about texture, density, and morphology. The features are fed to a classifier for the detection or the classification of the nodule.

The CAD system helps the radiologists to improve the reading of the computed tomography (CT) scans; however, a significant number of nodules remain undetected if a low positive rate is desired. This forbids the use of the CAD system in reality [14, 15]. There are variations in shapes, sizes, and types of the nodules, and some are even varied in texture and density. These wide variations are sometimes not diagnosed by the CAD system if the algorithm is not sophisticated enough.

Recently, because of the success of deep learning in numerous applications, CAD systems are also utilizing deep learning [16]. End-to-end deep learning has brought success in many medical image processing applications [17]. The pulmonary nodule detection systems from CT scan images also used several deep learning architectures in recent days [18–20]. These systems outperformed the systems using hand-crafted features [21].

As new healthcare is shifting towards smart healthcare, the use of wireless communication and mobile computing has been increasing in a smart healthcare framework. Until today, 3G/4G/5G communication is successfully used [22–24]. In [22], a block chain-based security scheme was proposed. An automatic seizure detection system using a mobile framework was proposed in [23]. A deep learning-based network resource algorithm in 5G was proposed in [24]. Now, the paradigm is shifting towards beyond 5G/6G to provide low latency, high transfer rate, and accommodate many sensors [25]. Smart systems are becoming popular in many applications [26, 27]. One of the important aspects of smart healthcare is to have a component of cognitive computing. Cognitive computing can facilitate health monitoring, medicine prescription, and mental state recognition [28]. The emotion of a human can tell a lot about the state of a patient. Therefore, recognizing the correct emotion can help understand the situation of a patient.

In a smart healthcare, a patient's case can be analysed by multiple doctors from various physical locations. A lung CT scan image can be uploaded to a computer system that can be accessed by several registered doctors. The system can produce an output of correct segmentation of nodules, if any, and provide a decision whether the image belongs to normal, benign, or malignant.

In this paper, we proposed a convolutional neural network- (CNN-) based pulmonary nodule detection and classification system. The classification outputs either one the three classes: normal, low level malignant, and high level malignant. The performance of the proposed system is compared with some of the state-of-the-art related systems.

The paper is structured as follows: Section 2 briefly outlines some of the previous related works. Section 3 describes the proposed system for detecting and classifying pulmonary nodules. Section 4 delivers experimental results and discussions. Section 5 concludes the paper.

2. Related Work

Most of the previous works used the Lung Image Database Consortium (LIDC-IDRI) database [29]. In different works, various numbers of samples from the database were used based on a selection criterion. In this section, we mainly focus on the works that used the LIDC database; however, some other important works are also mentioned.

First, we mention the works that used hand-crafted features to detect pulmonary nodules. Wu et al. proposed a nodule classification system using textual and radiological features [30]. They used 13 GLCM textual features and 12 radiological features along with a back-propagation neural network. A total of 2217 CT slices were used, and an area under the receiver operative characteristics (ROC) curve of 0.91 was obtained.

Shape and texture-based features together with a genetic algorithm and a support vector machine (SVM) were proposed to detect nodules in [31]. Before extracting features, the samples were enhanced by a quality threshold and a region growing-based segmentation. 97.5% accuracy was obtained with 140 samples from the LIDC database.

Orozco et al. developed a lung nodule classification system using 19 GLCM features extracted from different subbands of the wavelet transform and the SVM classifier [32]. The accuracy of 82% was obtained using a subset of the LIDC database. Han et al. used 3D GLCM features and the SVM for the nodule classification and got the area under the ROC curve of 92.7% [33]. Phylogenetic diversity index and genetic algorithm-based nodule classification systems were proposed in [34]. A total of 1403 images from the LIDC database were used in the experiments, and 92.5% accuracy was achieved by the system.

Second, we mention the works on pulmonary lung nodule detection and classification using deep learning. Mainly, we focus on the papers from 2018 onwards. A 4-channel CNN-based system was proposed in [35]. In this system, the scan images were enhanced by a Frangi filter, and the learning was based on a multigroup criterion. The LIDC images were used, and a sensitivity of 80.1% was obtained.

A topology-based phylogenetic diversity index on CT scans was used with CNN in [18]. 1404 images from the LIDC database were used in the experiments, and an accuracy of 92.6% was obtained. The images consisted of 394 malignant and 1011 benign nodules. A fusion of classifications using the Adaboost back propagation neural network was used in [36]. Three different types of features were utilized. One set of features was GLCM features, the second set of features was Fourier shape features, and the third set of features was obtained from a CNN architecture. These three sets of features were learned by three neural networks and fused. 1972 sample images (648 malignant and 1323

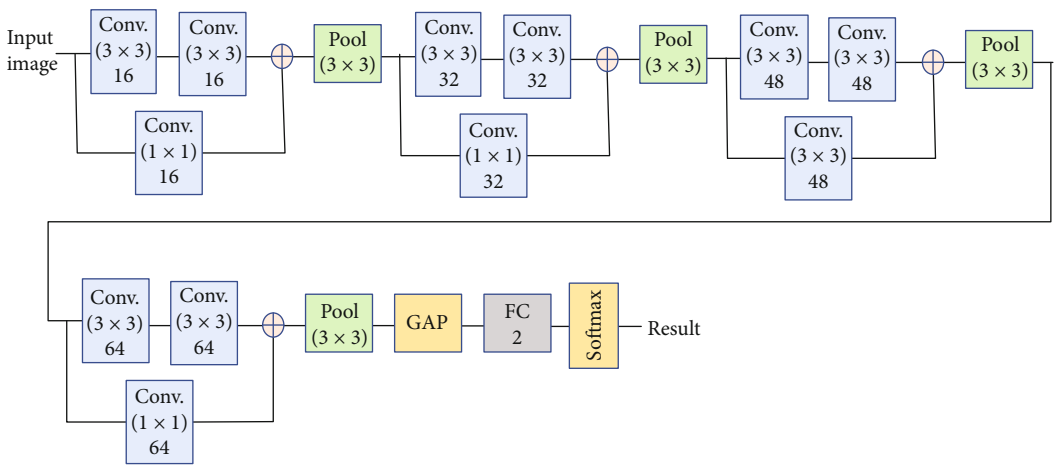


Figure 1: The architecture of the proposed light CNN model for lung nodule detection and classification.

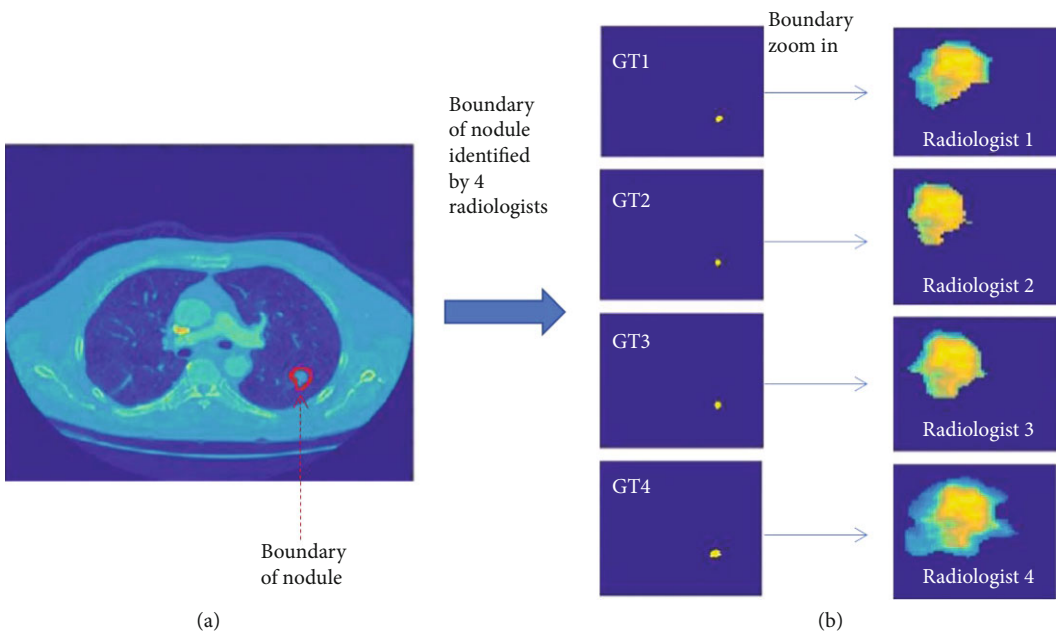


Figure 2: CT scan image (a), where the nodule is marked in the red circle. (b) shows the GTs and the NROIs determined by four radiologists.

benign) of the LIDC database were used in the experiments, and an accuracy of 96.7% was achieved.

Xie et al. proposed a nodule detection system using a faster region-based CNN [37]. The 2D convolutional operation was used to reduce false positives. The system achieved 86.4% accuracy using 150414 images. An end-to-end automated lung nodule detection system was developed in [38]. The system had three main phases. The system got 91.4% accuracy with false positives one per scan using 888 CT scans.

From the above review, we found that significant progress in lung nodule detection has been made during the last seven-eight years. The challenges are still there. The challenges include detection and classification of unevenly controlled nodules found on size, shape, and density. Therefore, there is a need for a fully automated system that can overcome some of these challenges.

3. Proposed System

Major CNN architectures such as AlexNet, VGG Net, and Google ResNet were designed to classify natural images that had around 1000 classes. These architectures were trained over millions of images and thereby were designed as very deep models. Medical structured data are not available in plenty, or the data size is limited. So, this limited data can cause overfitting in these architectures. Also, the visualization of medical data may not be meaningful using these very deep models.

3.1. CNN Architecture. In this paper, we developed a CNN architecture that is light (not very deep) and appropriate for medical image processing. The overall architecture is shown in Figure 1. There are four convolutional layers, followed by a global average pooling (GAP), two fully

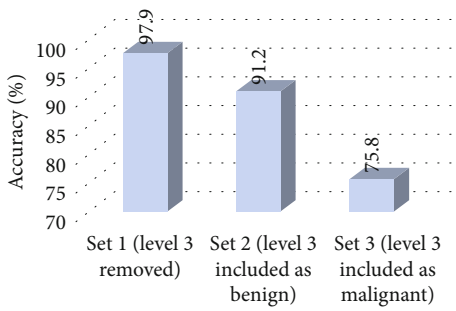


Figure 3: Accuracy (%) of the proposed system using three different sets of the LIDC database.

	Normal	Benign	Malignant
Normal	99.1%	0%	0.9%
Benign	0%	98.3%	1.7%
Malignant	1.0%	2.2%	96.8%

Figure 4: Confusion matrix.

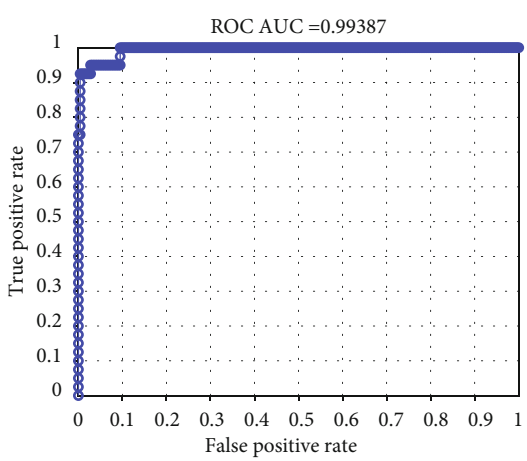


Figure 5: ROC curve of the system.

connected (FC) layers, and the softmax output layer, which has three output neurons corresponding to three classes (normal, benign, and malignant). Each convolutional layer has two successive 3×3 convolutional blocks with rectified linear units (ReLUs), a connector 1×1 convolutional block with the ReLU, and a max pooling block. In the first layer, the number of filters in each convolutional block is 16, for the second layer 32, for the third layer 48, and, finally, for the last layer 64. The stride of the filters is 1. The output of the connector convolutional block is summed with the output of the second CNN block before the max pooling. The stride of the max pooling is 2; so, the resolution is reduced by a factor of 4. Before each convolutional block, zero padding is applied to maintain the size. Mini batch normalization is applied to each layer to speed up the training. The GAP is used as a purpose of pooling, but it is more efficient than the pooling [39].

The input to the CNN is the image of size 256×256. The number of layers in the CNN is four so that the receptor window covers the whole image. We also tested with three layers; however, four layers performed better. Each pixel of the input image is normalized by the mean (mean subtraction) and standard deviation (divided by the standard deviation) of the pixels of the whole database.

The minibatch size was 4 samples, and the cost function was categorical cross-entropy. Before each minibatch, the samples were shuffled to ensure complete randomization of the learning; this also helped to overcome the overfitting. The initial weights were found by applying the normalization [40]. The Adam optimizer was used for optimizing the weights, and the parameters were $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1 \times 10^{-9}$, and the learning rate was 8×10^{-4}. The proposed CNN architecture is a modified version of the architecture proposed in [41]. The main difference between these two architectures is the number of layers; in our proposed architecture, we have a smaller number of layers, which makes the model a light model.

3.2. Database

3.2.1. Database Selection. The database that was used in the experiments is a publicly available database, named the LIDC-IDRI database [29]. There are 1018 CT scans of 1010 subjects from seven institutions. The slice thickness of the CT scans varied from 0.6 mm to 5.0 mm with a median of 2.0 mm. Four expert radiologists made the annotations of the scans in two separate reading sets. In the first set of readings, each suspicious lesion was classified independently as nonnodule, nodule with a size smaller than 3 mm, and nodule with size greater than or equal to 3 mm. In the second set of reading, 3D segmentation was done for the nodules which are greater than or equal to 3 mm.

3.2.2. Samples' Selection. The samples were selected in the experiments in the following manner. First, all the scans which had thickness above 3.0 mm were removed. Samples with nodule size less than 3 mm were also removed. Those nodules of size greater than or equal to 3 mm that were agreed by three or four radiologists were retained. The nodules were classified into different stages of malignancy and were ranked from malignancy level 1 to malignancy level 5. Levels 1 and 2 were denoted as benign, and levels 4 and 5 were denoted as malignancy. The samples with malignancy level 3 were not considered to make a clear distinction between benign and malignancy. Overall, there were 1279 samples selected for the experiments, of which 569 nonnodules, 278 benign, and 432 malignant. Figure 2 shows an example of a CT image, where the lung nodule is marked by

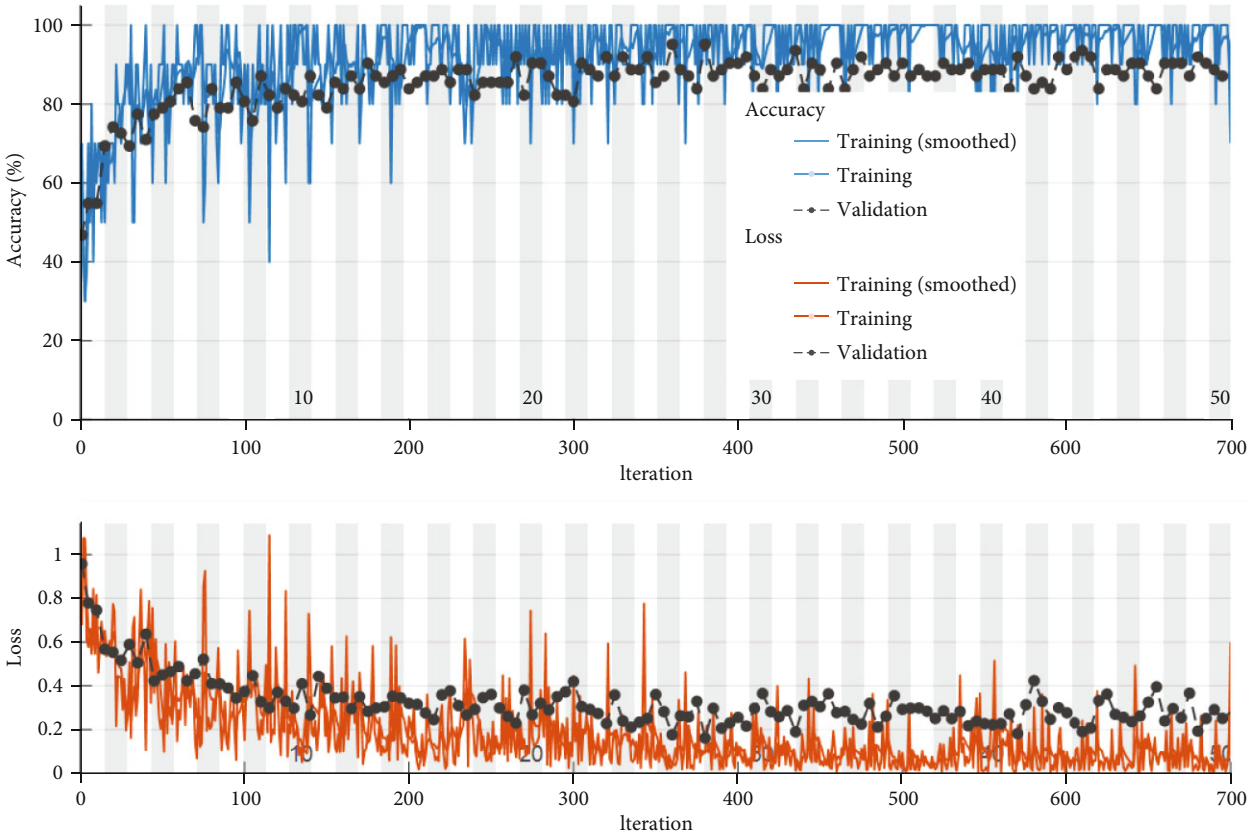


Figure 6: Learning curves (upper: accuracy; lower: loss) of the system.

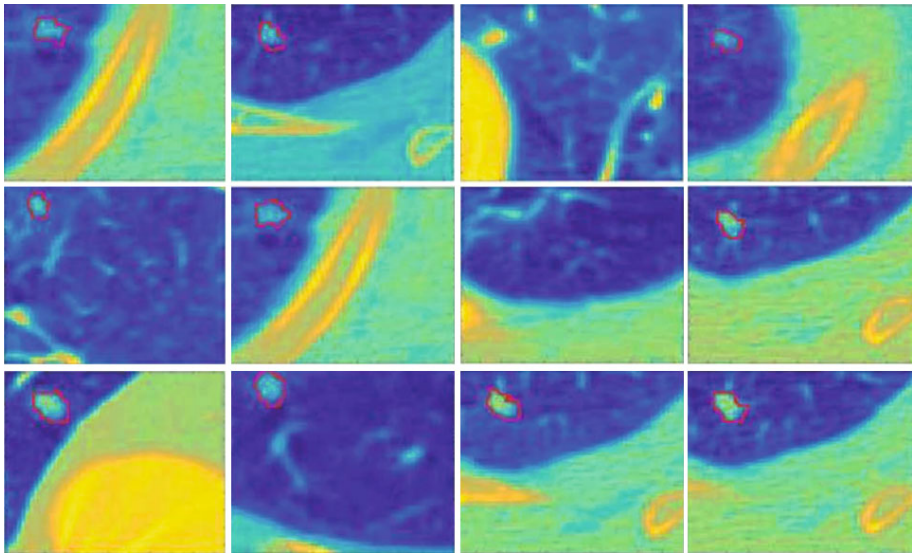


Figure 7: Some misclassified malignant samples, which were classified as benign samples.

a red circle. On the right side of the figure, there are ground truths (GTs) and corresponding segmentation as the nodule region of interest (NROI) by four radiologists. From the figure, we see that the radiologists' segmentations differ for a sample.

Nodule candidate regions are mined slice by slice from the LIDC. The candidate nodules' pixels retained their original values using a mask and make it a size 52×52 by padding zero as described in [41]. Eventually, all the samples are resized to 256×256.

3.2.3. Data Augmentation. The number of samples was not enough for proper training of the CNN, and also the

Table 1: Performance comparison between the systems.

System	Number of samples	Accuracy
[35]	1006 scans	Sensitivity = 80.6%
[18]	1011 benign, 394 malignant	92.6%
[36]	1011 benign, 394 malignant, 567 normal	96.7
Proposed	278 benign, 432 malignant, 569 normal	97.9%

numbers of the samples per class are unbalanced. Therefore, we need to raise the number of samples and balance the numbers by data augmentation. We applied the augmentation only for the training data. Only rotation and translation operations were used for the augmentation. The samples were rotated with random angles (between 10° and 60°) and translated within a range of [-2, 2].

4. Experimental Results and Discussion

The experiments were done by means of the 10-fold cross-validation approach. As described earlier, we removed level 3 samples to make a clear distinction between benign and malignant samples. In fact, in two sets of experiments, we also included level 3 samples. Therefore, we had three sets: set 1 had samples of level 3 removed, set 2 had samples of level 3 included in the benign category, and set 3 had samples of level 3 included in the malignant category. Set 1 had a total of 1279 samples of which 569 were normal, 278 were benign, and 432 were malignant. Set 2 had a total of 1508 samples, of which 507 were benign and 432 were malignant. Set 3 had 1508 samples of which 278 were benign and 661 were malignant. Figure 3 illustrates the accuracy of the proposed system using three sets. Set 1 had an accuracy of 94.65%, set 2 had 89.21%, and set 3 had 73.4%. From these results, we conclude that the samples of level 3 are more benign than malignant. In the subsequent experiments, we use only set 1.

Figure 4 displays the confusion matrix of the system using set 1. From the matrix, we find that the normal class generally was not confused with benign or malignant. Some benign and malignant samples were confusing between them. Malignant samples were confused the most.

We also found the confusion matrix recall and precision values of the system. The average recall was 98.07%, and the average precision was 98.06. Figure 5 shows the ROC curve of the system. The area under the curve was 0.987, which is considered very good. Figure 6 illustrates the learning curves in terms of accuracy and loss of the system. From the figure, we found that the accuracy and the loss are steady after some iterations. Figure 7 shows some malignant samples which were misclassified as benign samples. The misclassification samples did not have any specific criteria; however, the fading boundaries and size could contribute to such misclassification. We need more investigation into this matter.

Table 1 provides a measure of performance between systems. The proposed system was compared with some recent related systems which used deep learning. All the compared systems used the same LIDC database; however, the number of samples varied. The results of the systems were extracted from the corresponding papers. From the table, we find that the proposed system has got the highest accuracy. The closet accuracy was with the system in [36]. This system used three-streams and fused hand-crafted features with CNN features using an Adaboost neural network. Therefore, the system in [36] is heavily computationally intensive.

The proposed architecture has 275 MFLOPS and approximately 200 K parameters. On the other hand, the AlexNet has around 1.5 GFLOPS and 60 million parameters, and the GoogLeNet has around 3 GFLOPS and 7 million parameters. Therefore, our proposed architecture is very light compared to these famous architectures. All the experiments in this paper were carried out using a quad-core machine with 12 GB RAM and Nvidia GeForce GTX 1050 GPU.

5. Conclusion

The use of mobile computing, cognitive computing, machine learning, and healthcare data analytics greatly influences our life. To this end, a pulmonary nodule detection and classification system using a light CNN model was proposed. The system was evaluated using the LIDC database samples. The system achieved 97.9% accuracy when level 3 of malignancy samples was excluded in the experiments. The average recall and precision values were above 98%. Compared to the other state-of-the-art systems, the proposed system's performance is high. In a future study, we aim to visualize the nodule boundaries. We also want to fuse the features from different layers of the CNN architecture to enhance accuracy. Another direction is to use active learning to improve the performance [42].

Data Availability

Not applicable.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This study was funded by the Deanship of Scientific Research, Taif University, KSA (Research Project number 1-440-6146).

References

[1] Y. Zhang, M. S. Hossain, A. Ghoneim, and M. Guizani, "COCME: content-oriented caching on the mobile edge for wireless communications," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 26–31, 2019.

[2] J. Wang, Y. Miao, P. Zhou, M. S. Hossain, and S. M. M. Rahman, "A software defined network routing in wireless multi-hop network," *Journal of Network and Computer Applications*, vol. 85, pp. 76–83, 2017.

[3] K. Lin, C. Li, D. Tian, A. Ghoneim, M. S. Hossain, and S. U. Amin, "Artificial-intelligence-based data analytics for

cognitive communication in heterogeneous wireless networks," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 83–89, 2019.

[4] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive internet of things: improving sensitivity and interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58–64, 2019.

[5] G. Muhammad, M. F. Alhamid, M. Alsulaiman, and B. Gupta, "Edge computing with cloud for voice disorder assessment and treatment," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 60–65, 2018.

[6] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2399–2406, 2018.

[7] G. Muhammad, M. F. Alhamid, and X. Long, "Computing and processing on the edge: smart pathology detection for connected healthcare," *IEEE Network*, vol. 33, no. 6, pp. 44–49, 2019.

[8] A. Yassine, S. Singh, M. S. Hossain, and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing," *Future Generation Computer Systems*, vol. 91, pp. 563–573, 2019.

[9] M. Masud, M. S. Hossain, and A. Alamri, "Data interoperability and multimedia content management in e-health systems," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1015–1023, 2012.

[10] Z. Ali, G. Muhammad, and M. F. Alhamid, "An automatic health monitoring system for patients suffering from voice complications in smart cities," *IEEE Access*, vol. 5, no. 1, pp. 3900–3908, 2017.

[11] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2018.

[12] C. Jacobs, E. M. van Rikxoort, T. Twellmann et al., "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," *Medical Image Analysis*, vol. 18, no. 2, pp. 374–384, 2014.

[13] G. Xiuhua, S. Tao, W. Huan, and L. Zhigang, "Prediction models for malignant pulmonary nodules based-on texture features of CT image," in *Theory and Applications of CT Imaging and Analysis*pp. 63–76, IntechOpen, Noriyasu Homma.

[14] B. van Ginneken, S. G. Armato III, B. de Hoop et al., "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study," *Medical Image Analysis*, vol. 14, no. 6, pp. 707–722, 2010.

[15] M. Firmino, A. H. Morais, R. M. Mendoça, M. R. Dantas, H. R. Hekis, and R. Valentim, "Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects," *Biomedical Engineering Online*, vol. 13, no. 1, p. 41, 2014.

[16] M. S. Hossain and G. Muhammad, "Cloud-based collaborative media service framework for healthcare," *International Journal of Distributed Sensor Networks*, vol. 10, no. 3, Article ID 858712, 2014.

[17] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. Shamim Hossain, "Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Generation Computer Systems*, vol. 101, pp. 542–554, 2019.

[18] A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Classification of patterns of benignity and malignancy based on CT using topology-based phylogenetic diversity index and convolutional neural network," *Pattern Recognition*, vol. 81, pp. 200–212, 2018.

[19] N. Tajbakhsh and K. Suzuki, "Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: Mtanns vs. cnns," *Pattern Recognition*, vol. 63, pp. 476–486, 2017.

[20] X. Yuan, L. Xie, and M. Abouelenien, "A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data," *Pattern Recognition*, vol. 77, pp. 160–172, 2018.

[21] Y. Wang, Y. Qiu, T. Thai, K. Moore, H. Liu, and B. Zheng, "A two-step convolutional neural network-based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on CT images," *Computer Methods and Programs in Biomedicine*, vol. 144, pp. 97–104, 2017.

[22] M. A. Rahman, M. M. Rashid, M. S. Hossain, E. Hassanain, M. F. Alhamid, and M. Guizani, "Blockchain and IoT-based cognitive edge framework for sharing economy services in a smart city," *IEEE Access*, vol. 7, pp. 18611–18621, 2019.

[23] G. Muhammad, M. Masud, S. U. Amin, R. Alrobaea, and M. F. Alhamid, "Automatic seizure detection in a mobile multimedia framework," *IEEE ACCESS*, vol. 6, pp. 45372–45383, 2018.

[24] M. S. Hossain and G. Muhammad, "A deep-tree-model-based radio resource distribution for 5G networks," *IEEE Wireless Communications*, vol. 27, no. 1, pp. 62–67, 2020.

[25] Y. Zhang, Y. Qian, D. Wu, M. S. Hossain, A. Ghoneim, and M. Chen, "Emotion-aware multimedia systems security," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 617–624, 2019.

[26] A. Alelaiwi, A. Alghamdi, M. Shorfuzzaman, M. Rawashdeh, M. S. Hossain, and G. Muhammad, "Enhanced engineering education using smart class environment," *Computers in Human Behavior*, vol. 51, Part B, pp. 852–856, 2015.

[27] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying deep learning for epilepsy seizure detection and brain mapping visualization," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1s, pp. 1–17, 2019.

[28] Y. Hao, J. Yang, M. Chen, M. S. Hossain, and M. F. Alhamid, "Emotion-aware video QoE assessment via transfer learning," *IEEE MultiMedia*, vol. 26, no. 1, pp. 31–40, 2019.

[29] S. G. Armato III, G. McLennan, L. Bidaut et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[30] H. Wu, T. Sun, J. Wang et al., "Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography," *Journal of Digital Imaging*, vol. 26, no. 4, pp. 797–802, 2013.

[31] A. O. de Carvalho Filho, W. B. de Sampaio, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index," *Artificial Intelligence in Medicine*, vol. 60, no. 3, pp. 165–177, 2014.

[32] H. M. Orozco, O. O. V. Villegas, V. G. C. Sánchez, H. de Jesús Ochoa Domínguez, and M. de Jesús Nandayapa Alfaro, "Automated system for lung nodules classification based on wavelet

feature descriptor and support vector machine," *Biomedical Engineering Online*, vol. 14, no. 1, p. 9, 2015.

[33] F. Han, H. Wang, G. Zhang et al., "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *Journal of Digital Imaging*, vol. 28, no. 1, pp. 99–115, 2015.

[34] A. O. de Carvalho Filho, A. C. Silva, A. Cardoso de Paiva, R. A. Nunes, and M. Gattass, "Computer-aided diagnosis of lung nodules in computed tomography by using phylogenetic diversity, genetic algorithm, and svm," *Journal of Digital Imaging*, vol. 30, no. 6, pp. 812–822, 2017.

[35] H. Jiang, H. Ma, W. Qian et al., "An automatic detection system of lung nodule based on multigroup patch-based deep learning network," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1227–1237, 2018.

[36] Y. Xie, J. Zhang, Y. Xia, M. Fulham, and Y. Zhang, "Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT," *Information Fusion*, vol. 42, pp. 102–110, 2018.

[37] H. Xie, D. Yang, N. Sun, Z. Chen, and Y. Zhang, "Automated pulmonary nodule detection in ct images using deep convolutional neural networks," *Pattern Recognition*, vol. 85, pp. 109–119, 2019.

[38] X. Huang, W. Sun, T.-L. B. Tseng, C. Li, and W. Qian, "Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks," *Computerized Medical Imaging and Graphics*, vol. 74, pp. 25–36, 2019.

[39] A. A. Amory, G. Muhammad, and H. Mathkour, "Deep convolutional tree networks," *Future Generation Computer Systems*, vol. 101, pp. 152–168, 2019.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, Santiago, 2015.

[41] T. A. Lampert, A. Stumpf, and P. Gancarski, "An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2557–2572, 2016.

[42] G. Muhammad and M. F. Alhamid, "User emotion recognition from a larger pool of social network data using active learning," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10881–10892, 2017.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8848553, 11 pages
https://doi.org/10.1155/2020/8848553

Research Article

A Semi-Fragile Video Watermarking Algorithm Based on H.264/AVC

Chen Li, Yi Yang, Kai Liu, and Lihua Tian

School of Software Engineering, Xi'an Jiaotong University, Xi'an, China

Correspondence should be addressed to Lihua Tian; lhtian@xjtu.edu.cn

Received 27 March 2020; Revised 8 May 2020; Accepted 26 May 2020; Published 20 June 2020

Academic Editor: Huimin Lu

With the increasing application of advanced video coding (H.264/AVC) in the multimedia field, a great significance to research in video watermarking based on this video compression standard has been established. We propose a semifragile video watermarking algorithm, which can simultaneously implement frame attack and video tamper detection, herein. In this paper, the frame number is selected as the watermark information, and the relationship of the discrete cosine transform (DCT) nonzero coefficients is used as the authentication code. The 4×4 subblocks, whose DCT nonzero coefficients are sufficiently complex, are selected to embed the watermark. The parities of these nonzero coefficients in the medium frequency are modulated to embed watermarks. The experimental results show that the visual quality of the embedded watermarked video is virtually unaffected, and the algorithm exhibits good robustness. Furthermore, the algorithm can correctly implement frame attack and video tamper detection.

1. Introduction

The rapid development of internet and multimedia technology has brought about the increasing popularity of digital videos, such as network TVs, online videos, and mobile videos [1–5]. These video applications greatly enrich people's lives and engender ample convenience. However, the proliferation of videos and video-sharing also causes a variety of serious problems. With the help of various video editing tools, people can easily tamper and edit public video content illegally, as well as copy and spread it freely. However, these behaviors endanger the legitimate rights of the copyright owners. Recently, several major video hosting portals, such as Youku and Tudou, have been involved in copyright disputes, causing enormous losses to media manufacturers and restricting the application of digital multimedia. To protect the copyright ownership of original videos or to provide the content authentication of original works has become an urgent problem to grapple with. As an effective method to resolve the problem of copyright protection and content authentication of multimedia, digital watermarking technology has become a research focus in the field of information security [6–8]. The basic idea of video watermarking is to add some extra information to the original video without affecting its visibility, which can provide the evidence of copyright or integrity authentication for the video content when necessary.

With the acquirement of video quality, video sizes are becoming progressively larger. Due to their storage capacity and bandwidth, nearly all digital videos are transmitted with compression coding on the Internet or other transmission channels. Thus, video watermarking algorithms combining video compression standards are more applicable. As a new generation of video coding standard, the application of the advanced video coding (H.264/AVC) is widespread because of its higher compression efficiency and better network affinity. Presently, research on video watermarking algorithms based on H.264/AVC is becoming increasingly more active.

Usually, video watermarking is categorized into three types, namely, the fragile, robust, and semifragile watermarks. The fragile watermarking scheme is used to validate integrity authentication, and thus, it should be sensitive for all video manipulations with good transparency and large watermark capacity. Robust watermarking needs to resist most common video processing activities, such as recompression and filtering, and perhaps presents a greater sacrifice

on transparency and watermark capacity. It is mainly applied for copyright protection. Semifragile watermarking is insensitive to common video processing operations, but it is sensitive to malicious attacks, and thus, it is mainly applied in tamper detection.

These different types of watermarking schemes based on (H.264/AVC) have developed in varying degrees in recent years, i.e., all types of watermarking schemes based on H.264/AVC have undergone considerable development. More researchers are beginning to concentrate on the semifragile video watermarking method because of its application for tampering detection [9–14]. Among them, the proposed algorithms in [9–11] can detect and locate tampering with different precision levels. Farfoura [10] provides a semifragile watermarking scheme for H.264/AVC. The scheme has a certain robustness for content-preserving manipulations and is sensitive to manipulations with content-changing. The algorithm utilizes dual watermarking. One watermark is embedded into I-frames to detect spatial tampering, and the other watermark is embedded into P-frames to identify temporal attack. Xu et al. [13] proposed a novel semifragile watermark, which generates feature codes through block energy and then embeds watermarks by using discrete cosine transform (DCT) coefficients on the diagonal, which has little effect on video quality. Cedillo [14] calculated the variance of the original pixel and the reference pixel in the 4×4 blocks and selected the subblock with the smallest variance as the block to be embedded, and then hid the watermark content according to the coding characteristics of the prediction mode value. The algorithm has good transparency, but it needs to calculate the variance value of 4×4 blocks, which has a large computational complexity. In [9], a semifragile video watermarking algorithm is proposed. In the context-adaptive variable-length coding (CAVLC) entropy encoding process of luminance 4×4 block DCT coefficients, the watermark is embedded by modifying the last nonzero coefficient after zig-zag scanning. However, the algorithm is more complicated in design. In [15], a chaotic semifragile watermarking algorithm for copyright authentication is proposed. The time information of the video frame is used as a parameter combined with the modulation of the chaotic algorithm to be embedded in the video, thereby generating embedded watermark information and embedding it into the DCT coefficient. During the parsing process, mismatch of time information can be used to reveal tampering in the time domain. The disadvantage is that it is an uncompressed domain algorithm and cannot be applied to H.264/AVC. Facciol and Farrugia [16] proposed a reversible watermarking algorithm that applies differential extension rules to 4×4 DCT block quantized coefficients. The algorithm firstly performs 16 quantized 4 s based on zig-zag scanning at the encoding end. The 4×4 DCT coefficients are divided into eight groups, and it is determined whether or not each set of coefficients can be embedded in the watermark, and then the watermark is embedded therein.

Combining the advantages and disadvantages of the abovementioned existing algorithms, we herein propose an algorithm for simultaneously realizing video time domain and spatial domain integrity authentication. In our algorithm, the frame number is utilized as the watermark information, and the numerical relationship of the DCT coefficients is adopted as the authentication code. The watermark is embedded by changing the DCT coefficients of the luminance subblocks with more nonzero coefficients. Finally, the integrity of the video is detected by the watermark information, and the tamper detection is implemented by the authentication code.

This paper is organized as follows: in Section 2, the generation of watermarking and authentication code is given first, and then two parts of the scheme are briefly described in Section 3. Experimental results are given in Section 4, and the conclusions are drawn in the last section.

2. Generation of Watermark and Authentication Code

2.1. Generation of Watermark. To identify and verify a frame attack, one idea is to embed the frame number as watermark information into the current frame, such that the watermark information extracted from the video represents the frame number information of the video frame. If the extracted watermark can match the frame number, it indicates that the video is not subject to frame attack. Otherwise, the frame attack can be determined according to the relationship between the specific watermark and the frame number.

Here, we propose a new semifragile video watermarking algorithm, which is intended to encode the number of video frames into a binary sequence, and embed this binary sequence as watermark information into the current frame. After the watermark is extracted, the watermark sequence is decoded to a decimal number again. Finally, this value is compared with the number of the video frame to verify and identify the frame attack.

When the video is recompressed, the watermark extraction would be misplaced due to the transition of the prediction mode. A large number of experimental results prove that the higher the texture complexity of the video, the better the robustness of the watermark. To improve the robustness of the watermark, we choose to embed the watermark into the video region with a complex texture. To improve the correctness of frame number restoration, the watermark sequence is divided into three segments and embedded in the video in a loop. The specific watermark information scheme is as follows:

Step 1: combine the number of frames of a common video; we choose to encode the number value of the video frame into an 18-bit binary sequence. For example, if the frame number is 10, the binary sequence converted is 000000000000001010. The sequence does not have to be 18-bit, this can be decided according to the actual situation.

Step 2: divide the binary sequence into three groups. The first six values are the first group, the middle six values are the second group, and the last six values are the third group. We will embed these three sets of binary sequences as the watermark values to the frame. For example, if the frame number is 10, the binary sequence is 000000000000001010, the first set of sequences is 000000, the second set of sequences is 000000, and the third set of sequences is 001010.

Step 3: the watermark information sequence is handled by Arnold scrambles to obtain a secure watermark sequence. Finally, the watermark sequence should be embedded in the video repeatedly.

2.2. Generation of Authentication Code. The attack identification and verification of the frame can be implemented by the watermark information, and the tamper detection of the video requires the participation of the authentication code. The type of attack that the video is subjected to is usually determined based on the degree of change of the authentication code. Therefore, the authentication code needs to be insensitive to conventional attacks but sensitive to malicious attacks.

The main goal of major video tampering is to alter the interesting targets, which normally have complicated textures rather than the background, which is the flat region [17]. Therefore, we generate the authentication code from the region with complex texture region according to the DCT coefficients. In our algorithm, we select the numerical relationship of the video DCT coefficients as the authentication code, and the experimental results show that the authentication code is not sensitive to conventional attacks but to malicious attacks. The algorithm chooses to embed a watermark on the last nonzero coefficient. Therefore, the numerical relationship between the second last nonzero coefficient and the third last nonzero coefficient is specifically selected as the authentication code of the algorithm. The specific authentication code scheme is as follows:

Step 1: traverse each macroblock of each frame of video to determine the prediction mode of the macroblock. If the prediction mode of the macroblock is 16×16, skip it. If the prediction mode of the macroblock is 4×4, the execution continues.

Step 2: determine whether the macroblock includes six or more subblocks, which have at least three nonzero coefficients, if not, skip it. Otherwise, the execution continues.

Step 3: for the macroblock that satisfies the abovestated conditions, six subblocks with the most nonzero coefficients are selected, and the authentication code R is extracted in the order of the number of nonzero coefficients. Each subblock extracts a 1-bit authentication code, and each macroblock extracts a 6-bit authentication code. The authentication information is generated by comparing the relationship between the second last nonzero coefficient, $m1$, and the third last nonzero coefficient, $m2$, in each subblock, and the authentication codes are saved. The specific scheme is shown in Equation (1).

$$R = \begin{cases} 1 & \text{if } m1 \geq m2 \\ 0 & \text{else} \end{cases} \tag{1}$$

Step 4: after the aforementioned steps, we would get the authentication code array A[i]($i = 1, \cdots, M$, M is the sum of the authentication codes) and the authentication code flag array flag [j], corresponding to each 4×4 macroblock ($j = 1, \cdots, P$, P is the sum of all 4×4 macroblocks). Meanwhile, the watermarking image, including logo information are scrambled and converted into a binary sequence, W, through dimension reduction. Then, the length of the sequence, W, is calculated. Perform exclusive or operations on each element of array A and the corresponding element of array W in sequence, and the result is saved in array B.

3. Watermark Embedding and Extraction Algorithm

3.1. Watermark Embedding Algorithm. Different from robust video watermarking, a semifragile video watermark is often derived from the characteristics of the video itself. To avoid misplacing the extracted watermark, the macroblock that does not satisfy the certain condition can also be embedded with a watermark with a value of -1 to achieve the purpose of "placeholder." We choose to modify the parity of nonzero coefficients in a complex 4×4 macroblock to embed the watermark. The specific watermark embedding scheme is as shown in Figure 1, and the detailed steps are as follows:

Step 1: the watermark binary sequence $M = \{wm1, wm2, \cdots, wm18\}$ is equally divided into three sets of binary sequences.

Step 2: traverse all the macroblocks in the current frame, and select a macroblock that satisfies the following condition to perform the embedding of the binary watermark. The complex 4×4 macroblocks include six or more subblocks, which have at least three nonzero coefficients. To avoid nonsynchronization of the extracted watermark, all other macroblocks in the frame that do not satisfy the abovementioned conditions are embedded into six watermark information units with a value of -1, and the video content is not modified.

Step 3: embed the watermark in the macroblock that meets the requirements of Step 2. Next, embed six binary watermarks in each luma macroblock, and select the watermark sequence value to be embedded, according to the remainder of the macroblock number divided by three.

Step 4: for the macroblock that satisfies the abovestated conditions, select six subblocks with the most nonzero coefficients, and embed the watermark in the order of the number of nonzero coefficients.

Step 5: for each filtered subblock, square the sum of the second last nonzero coefficient, $m1$, and the third last nonzero coefficient, $m2$, in each subblock. The watermark is embedded by modifying the last nonzero coefficient to satisfy the identical parity of the sum and the watermark value. The specific modification approach is as follows:

if $wm = 0$ and $LNZ^2 > sum$:

$$LNZ = \begin{cases} \lfloor \sqrt{sum} \rfloor & LNZ \geq 0 \\ -\lceil \sqrt{sum} \rceil & LNZ < 0 \end{cases} \tag{2}$$

if $wm = 1$ and $LNZ^2 \leq sum$:

$$LNZ = \begin{cases} \lfloor \sqrt{sum} \rfloor & LNZ \geq 0 \\ -\lceil \sqrt{sum} \rceil & LNZ < 0 \end{cases} \tag{3}$$

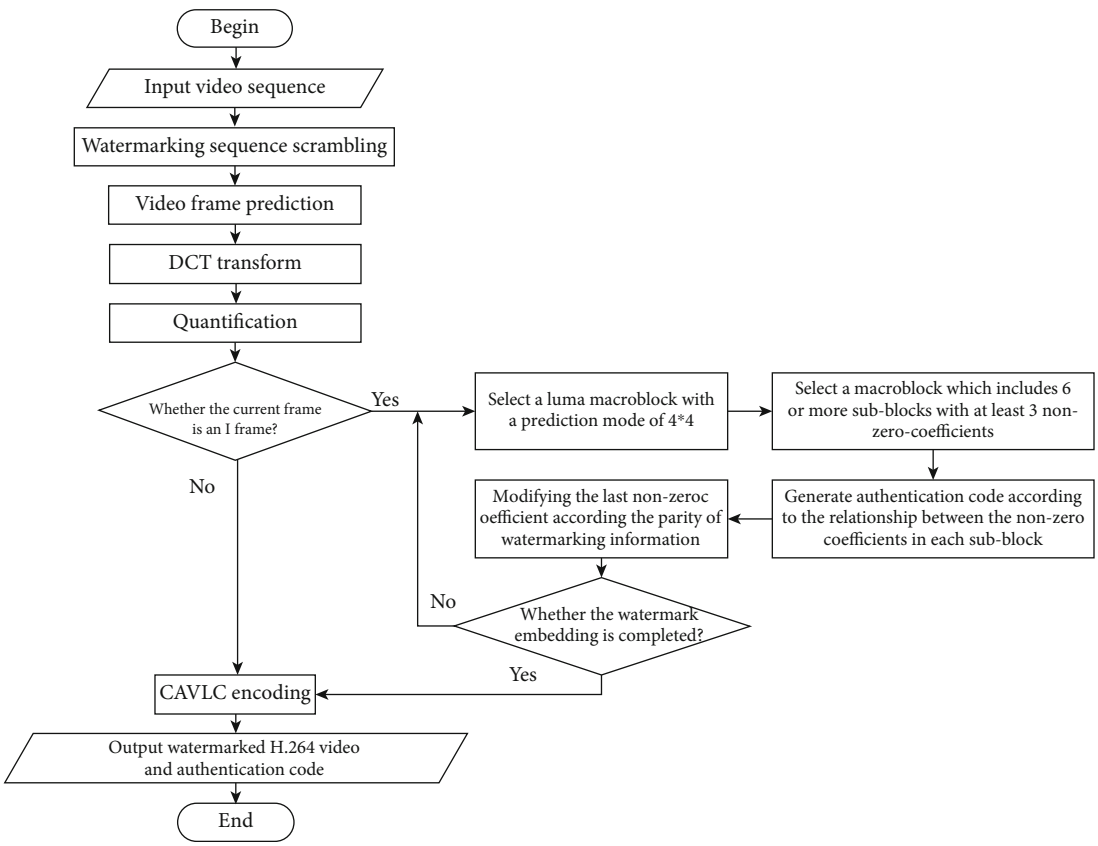


Figure 1: Flow chat of watermark embedding.

where LNZ is the last nonzero coefficient in NNZ, sum is the sum of all nonzero coefficients, wm is the embedded watermark, and $\lfloor\rfloor, \lceil\rceil$ represent rounding up and rounding down.

3.2. Watermark Extraction Algorithm and Authentication Code Judgment. The process of watermark extraction is equivalent to the inverse process of watermark embedding. The extracted watermark information contains the information of the frame number, so the frame number needs to be restored. The specific watermark extraction scheme is shown as Figure 2, and the detailed steps are as follows:

Step 1: traverse all the macroblocks in the current frame, and select a macroblock that satisfies the following condition to perform the embedding of the binary watermark. The prediction mode is 4×4 and includes six or more subblocks, which have at least three nonzero coefficients. For all macroblocks in the frame that do not satisfy the abovementioned conditions, extract six watermarks with a value of -1.

Step 2: if the macroblock satisfies the aforementioned conditions, the six subblocks with the most nonzero coefficients are selected, and the watermark is extracted according to the order of the number of nonzero coefficients. If the square of the last nonzero coefficient is larger than the square sum of the second last and third last nonzero coefficients, then the corresponding watermark is one. If the square of the last nonzero coefficient is less than or equal to the square sum of the second last and third last nonzero coefficients, then the corresponding watermark value is zero. At the same time, the authentication code is restored.

Step 3: after extracting all the watermarks, the watermark sequences are grouped and classified. The watermark sequence should be classified into three groups. If the macroblock number is divided by three and the remainder is one, the watermark sequence corresponding to the macroblock belongs to the first group, and the other types are the same. The most frequently occurring watermark sequence is the correct result. The three sets of sequence values are then integrated into an 18-bit binary sequence that is converted to a decimal number, which is the number of the current frame.

Here, we should first determine whether the whole video has been maliciously attacked by the extracted watermark information. If the watermark information is completely correct, we assume that the video has not been tampered maliciously. Otherwise, the correct frame number cannot be extracted from the watermark sequence, and we presume that the video may be attacked unconventionally. In this case, the authentication code is adopted to determine the occurrence of tampering and locate the specific modification area in the tampered frames.

Given that we make use of the relationship between the second last nonzero coefficient, $m1$, and the third last nonzero coefficient, $m2$, as the authentication code, recompression can be resisted. If malicious content tamper in a frame occurs, it can be recognized because there would be a large

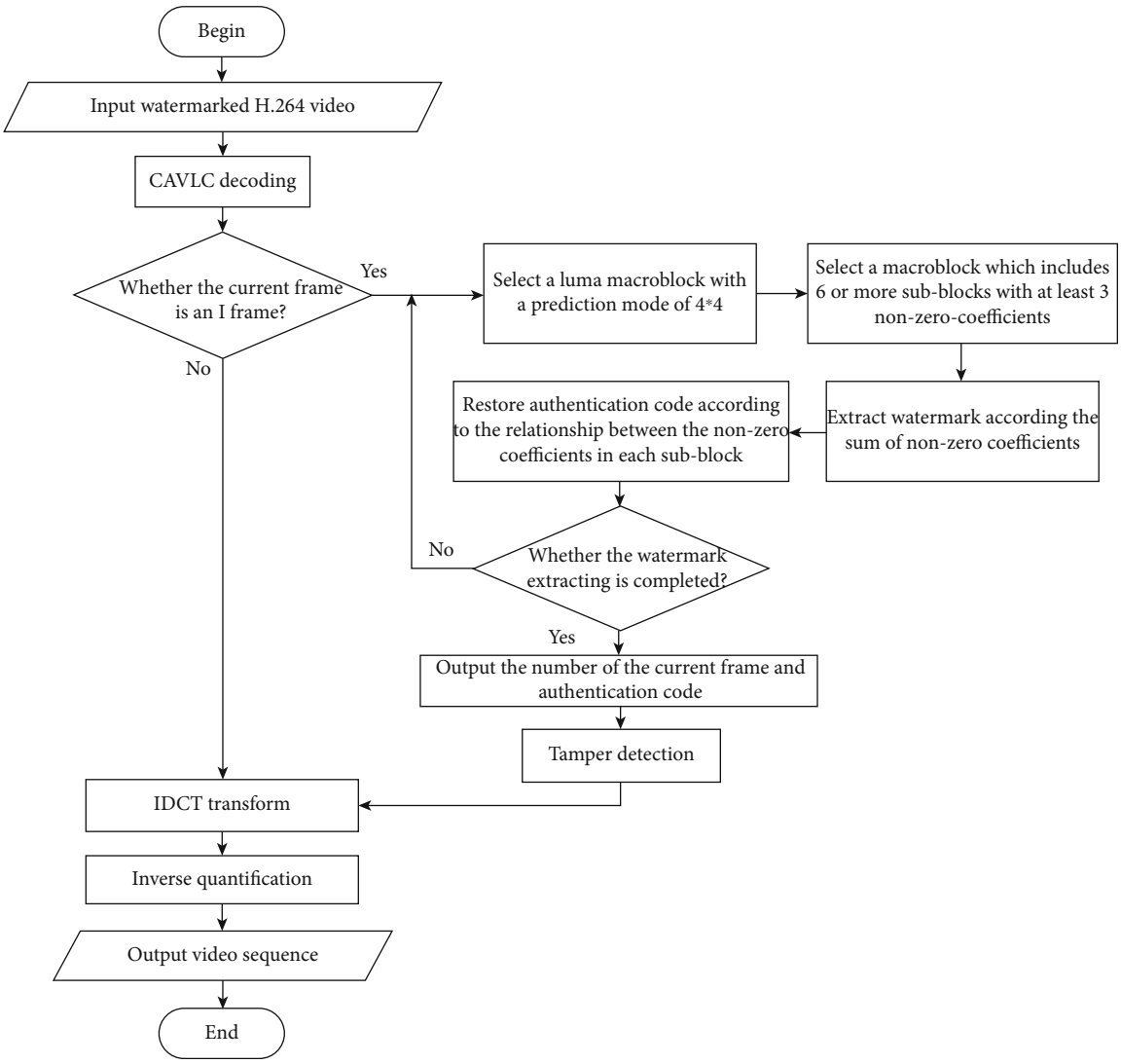


Figure 2: Flow chat of watermark extraction.

change in the authentication code. The specific criteria for judgment are as follows:

Step 1: firstly, check whether the extracted watermark has problems in the macroblock. If the watermark information does not synchronize, continue to Step 2. Otherwise, we consider that there is no tampering.

Step 2: if the frame number based on the watermark extraction is incorrect, it is indicative that the video has been subjected to conventional processing or malicious tampering. Firstly, we should judge whether the flag of this macroblock is variant. If the flag is changed, the extracted information of the current macroblock is distant from the information in the encoder, and we can directly affirm that the current macroblock has been tampered. Otherwise, continue to Step 3.

Step 3: under the condition that the flag is changeless and the watermark is extracted successfully, we generate the authentication code through the same method in encoding. The number of 6-bit authentication code changed is necessary to make further decisions.

Figure 3: Original logo image.

If there are K or more authentication codes that change greatly in the current macroblock, it means that the macroblock has been tampered with maliciously. Otherwise, it means that the macroblock has been subjected to conventional video processing. Here, K is a threshold, and it should be adjusted by the complexity of the video. Denote the authentication code sequence in the decoding as array C, and perform exclusive or operations on each element of array A and the corresponding element of array C in sequence, and the result is saved as W'. Rearranging W' into a two-dimensional array, the logo image can be reconstructed.

(a) Original image (b) Watermarked image (c) Pixel difference

Figure 4: Video frame before embedding watermark and after embedding watermark.

4. Results and Discussion

In this section, we would test the performance of our algorithm under different conditions. We realize our algorithm based on the reference joint test model (JM) software JM by H.264 using the version 8.6. Standard video sequences, including akiyo_qcif, container_qcif, mobile_qcif, news_qcif, tempete_qcif, carphone_qcif, coastguard_qcif, silent_cif, highway_cif, and flower_cif, are provided to demonstrate the effectiveness of the proposed watermarking algorithm. The size of all videos are 176×144 and 352×288 with quarter common intermediate format (QCIF) and common intermediate format (CIF), which adopts the baseline encoding mode. The frame rate is set to 30 frames per second, and the default quantization parameter is set to 28. The logo image is 32×32 two-valued images, as shown as Figure 3.

4.1. Imperceptibility Test. To test the imperceptibility, we provide the comparison of the original sequence and the same reconstructed frame with embedded watermark for the different video. By the space limitation, we provide three video results, which are shown in Figure 4, whose left column is some frame of original sequence, and middle column is the corresponding frame of the same sequence with watermarking, and the right column is the pixel difference of both as seen in Figure 4; the video quality is almost unchanged by subjective judgment.

To better evaluate the invisibility of our watermarking method in an objective way, we introduce two evaluation standards of video quality evaluation, i.e., PSNR (peak signal to noise ratio) and SSIM (structural similarity index) [18] to compare the original video and the video with the embedded watermark. To eliminate the influence of randomness, we give the average of the PSNR and SSIM for all the frames in the video and compare our algorithm to [10]. The comparison results are as shown in Figures 5 and 6, respectively. Normally, if the PSNR is higher than 30, the quality of the processed image is better. From Figure 5, we can see that the PSNRs of our algorithm for various formats and different videos are all higher than 34.0, which proves that our method has inconsiderable effect on video quality. We also compare to [19], the result is shown in Figure 6. Though our method is slightly inferior than [19], the difference is very small, and other indicators have greater ground. From Figure 7, we can know that all the SSIMs of our algorithm are above 0.960, which is very close to 1, and this also proves that the image quality is fine. All SSIM results of different videos are better than [10], as shown in Figure 8. Because we only embed the information into the areas with more complex textures, the impact it has on the video quality is minimal. The abovementioned experiment results evidently prove this fact.

4.2. Bit Change Rate Test. For a watermarking method, the bit-change rate should be also considered. After all, nobody likes the size of the compressed video becoming bigger after

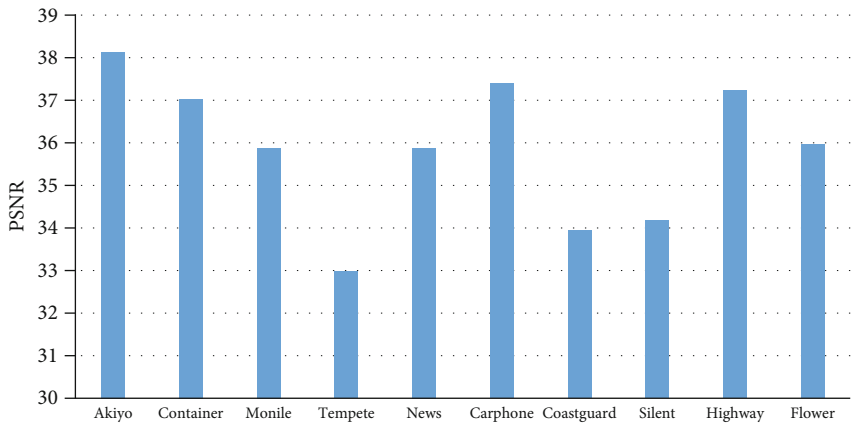


Figure 5: PSNR of proposed method for different sequences.

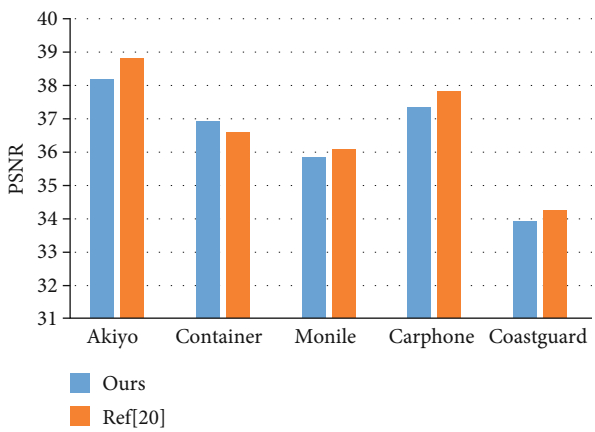


Figure 6: PSNR comparison for different video sequences.

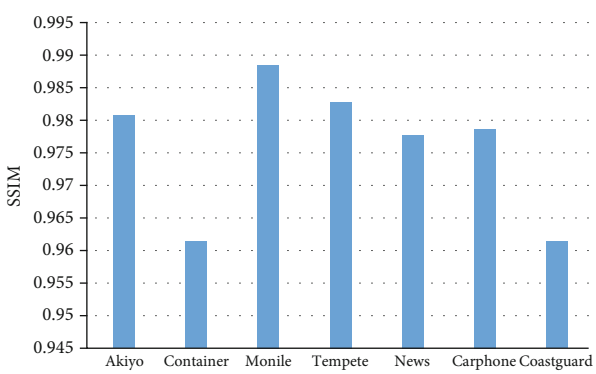


Figure 7: SSIM of proposed method for different sequences.

embedding the watermark. If the bit-change rate is smaller, we can get better video affinity [20]. We can calculate the bit-change rate as follows:

$$BIR = \frac{\left(BIR_{water}\text{-}BIR_{org}\right)}{\left(BIR_{org}\right)} \times 100 \quad (4)$$

The bit-change rate of our algorithm is very low, and the bit-increase rates of the video sequence of container, foreman, mobile, and news are only 0.08%, 0.06%, 0.04%, and 0.08%, respectively. Notwithstanding, the bit-rate changes in [17] are apparent. Comparing with [10, 17, 19], the results are shown in Figure 9. From the comparison results, we can see that our algorithm outperforms the compared ones. The smaller the bit-rate change, the better the real-time performance of the algorithm. Therefore, the proposed algorithm has a better real-time performance and is more suitable for real-time broadcast applications.

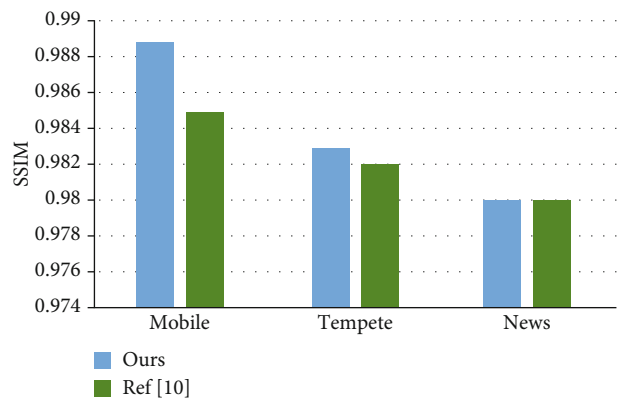


Figure 8: SSIM comparison for different video sequences.

4.3. Robustness Test. Several experiments are conducted to verify the robustness of our watermarking algorithm against recompression. At first, the extracted watermark is provided in Table 1, under the condition that the video with the embedded watermark information is recompressed with equal quantization values ($QP = 28$ and $QP = 34$). From Table 1, we can see that the logo image is clear, which reflects that our algorithm has better performance against recompression.

For objectively evaluating the performance of our algorithm against recompression, the NC (bit accuracy rate) and BER (bit error rate) are adopted to further measure the robustness of the algorithm. The closer the NC is to one and the nearer the BER is to zero, the video distortion from the watermark is less, and the robustness of the algorithm is better. The experimental results of the NC and BER are given

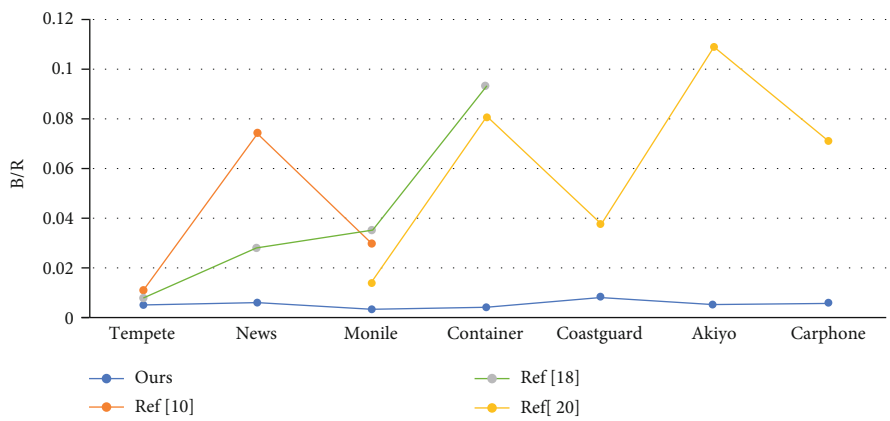
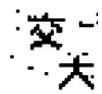
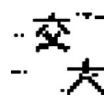
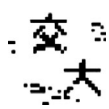
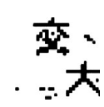
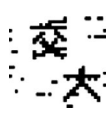


Figure 9: Bit rate contrast diagram.

Table 1: The extraction result under different QP.

Sequence	$QP = 28$	$QP = 34$
Akiyo		
Carphone		
Container		
Coastguard		
Mobile		
News		
Tempete		

in Figures 10 and 11, respectively. From them, it is revealed that under the condition of equal quantization and recompression, the NC is higher than 0.91, and it is stable at the mean value of 0.97. When the NC is higher than 0.8, it means that this algorithm has decent robustness. In the meantime, the BER of all sample videos is lower than 0.01 and the average BER is 0.028, which indicates that the proposed method has good robustness for recompression. Compared with [17], our algorithm has less BER than [10, 17] for the same test video, which proves the algorithm has stronger robustness under equal compression quantization, as shown in Figure 11.

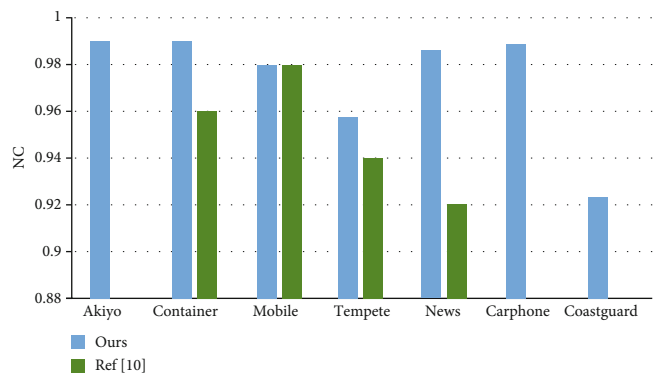


Figure 10: NC comparison under equal quantized recompression.

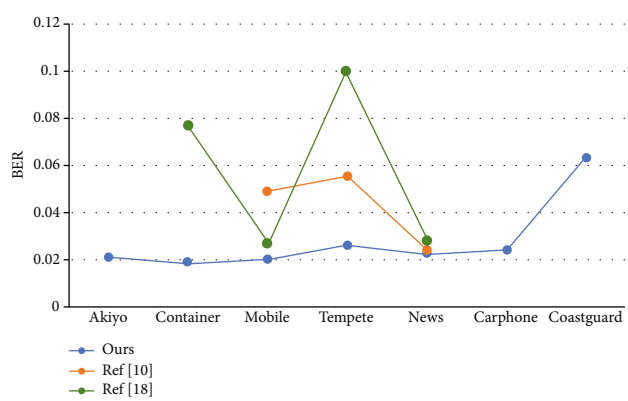


Figure 11: BER comparison under equal quantized recompression.

4.4. Tamper Detection Test. Based on our watermarking method, the authentication code is used to locate intraframe tampering, and the watermark is used to detect malicious operations, such as addition and deletion between frames. To test the performance of the algorithm to resist the interframe attacks, many experiments are designed, including frame dropping, frame addition, and frame replacement. In the experiment of the frame dropping attack, the frames 40–80 of the video sequence are dropped. The detected result of our method is shown in Figure 12. We can see clearly that our method can successfully extract the reserved frame

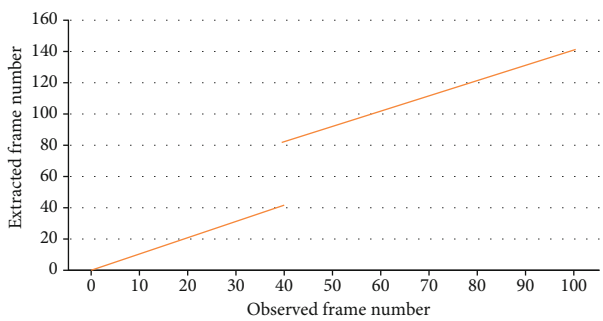


Figure 12: Temporal tampering frame dropping.

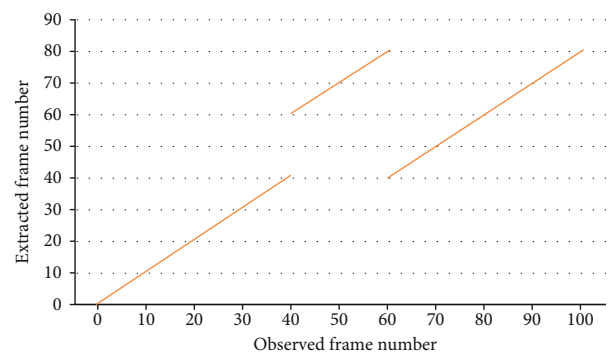


Figure 13: Temporal tampering frame addition.

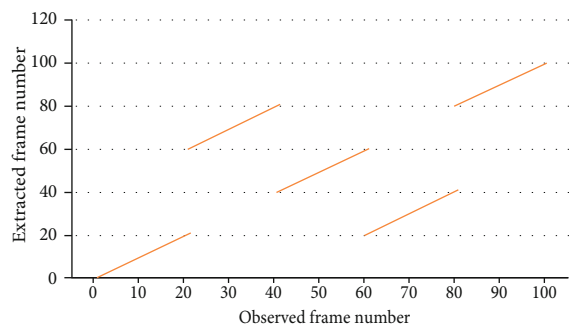


Figure 14: Temporal tampering frame replacing.

number (1–40) and identify the removal of frames 40–80. In the experiment of the frame adding attack, we added frames 61–80 to the 41st frame. The Figure 13 shows the results, wherein we can find our method is also effective for this attack. Finally, we experiment the method with frame replacing attack by substituting frames 61–80 with frames 21–40, and Figure 14 shows that our algorithm can detect such changes successfully.

Through experiments, we also provide the proportion of verification code changes of different video sequences after recompression. The statistical results are shown in Table 2 below.

Then, the proportion of verification code changes of different video sequences after tampering is also provided. The statistic results are shown in Table 3 below.

From the above tables, it is found that the rate of change of videos after recompression and tampering varies greatly.

Next, we tested the effectiveness of the proposed method to the tamper in a frame. Based on the method, if the current macroblock has K or more authentication codes changed, it indicates that the macroblock has been maliciously tampered. Otherwise, the macroblock is processed normally. The value of K is determined according to the actual situation. In this experiment, we implement the specific tamper to copy and paste the top right corner of the first frame of the mobile video sequence. By comparing the authentication codes of each macroblock whether changed to prove the effectiveness of the proposed method. The results of the experiment are shown in Figure 15. The right column of Figure 15 gives the location of the tamper, and we can see an evident region in the top right side.

The detection accuracy rate (DAR) is computed by $DAR = (TPR + TNR)/2$, where TPR and TNR denote true positive rate and true negative rate, respectively. Table 4 shows the experimental results that represent the DAR of our method on different QPs. Also, the results illustrate that our method is efficient to handle content based tamper detection.

Table 2: Rate of change of authentication code after recompression.

Sequence	QP	Rate of change
Akiyo	28	16.1%
Container	28	15.5%
Mobile	28	9.4%
News	28	19.3%
Tempete	28	17.6%
Carphone	28	19.6%
Coastguard	28	18.2%

Table 3: Rate of change of authentication code after tampering.

Sequence	QP	Rate of change
Akiyo	28	79.7%
Container	28	76.9%
Mobile	28	77.3%
News	28	82.5%
Tempete	28	80.7%
Carphone	28	75.6%
Coastguard	28	81.2%

5. Conclusions

Herein, we propose a semifragile video watermarking algorithm based on content authentication, which has good robustness to recompression under different QPs and can simultaneously implement frame attack and video tamper detection. The frame number is binary and is converted into an 18-bit long binary sequence as a watermark. The sequence is divided into three sets of NNZ residual coefficients embedded in the DCT of the video to detect frame attack in the video. To realize the tampering detection of the video, the

Figure 15: Tampering detection diagram.

Table 4: Accuracy evaluation of tamper detection.

Sequence	TPR	TNR	DAR
News	84.08%	84.24%	84.16%
Mobile	88.36%	87.68%	88.02%
Tempete	86.42%	85.32%	85.87%

size relationship of the DCT nonzero coefficient is selected as the authentication code to distinguish whether the video is subjected to normal operation or malicious tampering. The experimental results show that the algorithm has good watermark invisibility, and the algorithm uses a multivalued watermark to embed, which avoids the nonsynchronization of the extracted watermark and greatly reduces the BER value of the watermark after recompression. Compared with the current algorithms, the algorithm has better invisibility and robustness, and it also has a good ability of frame attack and video tamper detection.

Data Availability

The video we use to test can be downloaded from http://trace.eas.asu.edu/yuv/index.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant no. 61901356 and the HPC Platform of Xi'an Jiaotong University.

We would like to thank Editage (http://www.editage.com/) for English language editing.

References

[1] S. Serikawa and H. Lu, "Underwater image dehazing using joint trilateral filter," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 41–50, 2014.

[2] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain Intelligence: go beyond artificial intelligence," *Mobile Networks and Applications*, vol. 23, no. 2, pp. 368–375, 2018.

[3] H. Lu, Y. Li, T. Uemura, H. Kim, and S. Serikawa, "Low illumination underwater light field images reconstruction using deep convolutional neural networks," *Future Generation Computer Systems*, vol. 82, pp. 142–148, 2018.

[4] Y. Sakai, H. Lu, J. K. Tan, and H. Kim, "Recognition of surrounding environment from electric wheelchair videos based on modified YOLOv2," *Future Generation Computer Systems*, vol. 92, pp. 157–161, 2019.

[5] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2400–2413, 2020.

[6] S. Gaur and V. K. Srivastava, "A hybrid RDWT-DCT and SVD based digital image watermarking scheme using Arnold transform," *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2017, pp. 399–404, Noida, India, 2017.

[7] S. Dong, J. Li, and S. Liu, "Frequency domain digital watermark algorithm implemented in spatial domain based on correlation coefficient and quadratic DCT transform," in *2016 International Conference on Progress in Informatics and Computing (PIC)*, pp. 596–600, Shanghai, China, 2016.

[8] S. J. Horng, D. Rosiyadi, P. Fan, X. Wang, and M. K. Khan, "An adaptive watermarking scheme for e-government document images," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 3085–3103, 2014.

[9] X. L. Chen and H. M. Zhao, "A novel video content authentication algorithm combined semi-fragile watermarking with compressive sensing," in *2012 Second International Conference on Intelligent System Design and Engineering Application*, pp. 134–137, Sanya, Hainan, 2012.

[10] M. E. Farfoura, S.-J. Horng, J.-M. Guo, and A. Al-Haj, "Low complexity semi-fragile watermarking scheme for H.264/AVC authentication," *Multimedia Tools and Applications*, vol. 75, no. 13, pp. 7465–7493, 2016.

[11] A. K. Mairgiotis and D. Ventzas, "Video watermark detection in DCT domain for H.264/AVC and extensions through a hierarchical prior," *2nd IET International Conference on Intelligent Signal Processing 2015 (ISP)*, 2015, pp. 1–6, London, 2015.

[12] M. Fallahpour, S. Shirmohammadi, M. Semsarzadeh, and J. Zhao, "Tampering detection in compressed digital video using watermarking," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 5, pp. 1057–1072, 2014.

[13] D. Xu, R. Wang, and J. Wang, "A novel watermarking scheme for H.264/AVC video authentication," *Signal Processing: Image Communication*, vol. 26, no. 6, pp. 267–279, 2011.

[14] A. Cedillo-Hernandez, M. Cedillo-Hernandez, M. Garcia-Vazquez, M. Nakano-Miyatake, H. Perez-Meana, and A. Ramirez-Acosta, "Transcoding resilient video watermarking scheme based on spatio-temporal HVS and DCT," *Signal Processing*, vol. 97, pp. 40–54, 2014.

[15] S. Chen and H. Leung, "Chaotic watermarking for video authentication in surveillance applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 704–709, 2008.

[16] R. Facciol and R. Farrugia, "Robust Video Transmission using Reversible Watermarking Techniques," in *2010 IEEE International Symposium on Multimedia*, pp. 161–166, Taichung, Taiwan, 2010.

[17] L. Tian, H. Dai, and C. Li, "A semifragile video watermarking algorithm based on chromatic residual DCT," *Multimedia Tools and Applications*, vol. 79, no. 3-4, pp. 1759–1779, 2020.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] N. Mehmood and M. Mushtaq, "Blind watermarking scheme for H.264/AVC based on intra 4x4 prediction modes," in *Future Information Technology, Application, and Service*, pp. 1–7, Springer, 2012.

[20] L. Tian, N. Zheng, J. Xue, and T. Xu, "A CAVLC-based blind watermarking method for H. 264/AVC compressed video," in *2008 IEEE Asia-Pacific Services Computing Conference*, pp. 1295–1299, Yilan, Taiwan, 2008.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 2138643, 12 pages
https://doi.org/10.1155/2020/2138643

Research Article

Simultaneous Localization and Mapping Based on Kalman Filter and Extended Kalman Filter

Inam Ullah,[1] **Xin Su**,[1] **Xuewu Zhang**,[1] **and Dongmin Choi**[2]

[1]*College of Internet of Things (IoT) Engineering, Hohai University (HHU), Changzhou Campus, 213022, China*
[2]*Division of Undeclared Majors, Chosun University, Gwangju 61452, Republic of Korea*

Correspondence should be addressed to Dongmin Choi; jdmcc@chosun.ac.kr

Received 28 January 2020; Revised 14 March 2020; Accepted 22 May 2020; Published 8 June 2020

Academic Editor: Yin Zhang

For more than two decades, the issue of simultaneous localization and mapping (SLAM) has gained more attention from researchers and remains an influential topic in robotics. Currently, various algorithms of the mobile robot SLAM have been investigated. However, the probability-based mobile robot SLAM algorithm is often used in the unknown environment. In this paper, the authors proposed two main algorithms of localization. First is the linear Kalman Filter (KF) SLAM, which consists of five phases, such as (a) motionless robot with absolute measurement, (b) moving vehicle with absolute measurement, (c) motionless robot with relative measurement, (d) moving vehicle with relative measurement, and (e) moving vehicle with relative measurement while the robot location is not detected. The second localization algorithm is the SLAM with the Extended Kalman Filter (EKF). Finally, the proposed SLAM algorithms are tested by simulations to be efficient and viable. The simulation results show that the presented SLAM approaches can accurately locate the landmark and mobile robot.

1. Introduction

Wireless sensor networks (WSNs) grasp the potential of various new applications in the area of management and control. Examples of such applications include detection, target tracking, habitation monitoring, catastrophe management, and climate management such as temperature and humidity. The key technology that drives the development of sensor applications is the quick growth of digital circuit mixing. In the recent future, these applications will provide a small, cheap, and efficient sensor node. Localization is also crucial for various applications in WSNs. In both universal computing and WSNs, there has been considerable consideration of localization [1, 2]. Characteristically, the WSN system offers the range and/or bearing angle measurements between each landmark and vehicle.

Mobile robot localization is also one of the attractive researches that support a truly self-governing mobile robot performance. Various independently working robots can accomplish tasks more rapidly in many situations. However, there is a possibility of even better productivity gains if robots can work cooperatively. The capability to collaborate is dependent on the robot's capability to connect and communicate with each other's. In this work, the authors consider the procedure of simultaneous localization and mapping (SLAM). For this purpose, a linear Kalman Filter (KF) with SLAM and Extended Kalman Filter (EKF) with SLAM are applied [3, 4].

SLAM plays a key role in the field of robotics and especially in a mobile robot system. The key objective of SLAM is to jointly measure the position of the robot as well as the model of the surrounding map [5–7]. For the safe interaction of robots within the operation area, this information is important. A variety of the SLAM algorithm has been presented over the last decade. Most of the early algorithms for SLAM used a laser rangefinder [8] which works as the core sensor node, and visual sensor nodes are the most used option currently, whichever is active or passive [9, 10]. In contrast to a laser rangefinder, currently, small, light, and affordable cameras can offer higher determination data and virtually unrestricted estimation series. These cameras work as passive sensor nodes and, therefore, do not affect one

another while deploying in similar operation areas. Distinct in the designed light range sensor nodes, cameras are also able to apply for both interior and exterior situations. Therefore, such features can make the camera the best choice for mobile robotic platforms and SLAM.

In SLAM, the need for using the environment map is twofold or double [11, 12]. The first one is the map often essential to support or back up other responsibilities; for example, a map can notify a track arrangement or offer an initiative imagining for a worker. Secondly, the map or plot follows in restraining the fault performed in measuring the state of the robot. While without a map, the dead reckoning would rapidly point energetically. On the other hand, by using a map, for example, a set of distinct landmarks, the robot can reorganize its localization error by reentering the known areas. Therefore, SLAM applications are more useful in such situations in which a preceding plan is not existing and require to be constructed. In some aspects of the robots, a set of landmark location is known prior. For example, a robot is operational on the floor of a workshop that can be supplied with a physically assembled chart of artificial guidelines in the operation area. Alternatively, in another case, in which the robot has admittance to the global positioning system (GPS), the GPS satellite can be chosen as a moving beacon at a prior known position. In this case, the SLAM may not be needed if the localization is done consistently concerning the prior known landmark of the robot. Through the development of indoor localization uses of mobile robots, the popularity of SLAM is increased. Most of the indoor procedures rule out the practice of GPS to assure the error of localization. The SLAM algorithm also provides an interesting substitute to the maps which is built by the user, which represents that the process of the robot is also conceivable in the nonappearance of ad hoc networks for localization [13].

KF derivatives are concerned with the first branch of those methods which apply a filter [14, 15]. The KFs assume that Gaussian noises affect data, which is not inevitably accurate in our case. KFs are planned to solve the problems of linear systems in their basic form and are rarely used for SLAM, although they have great convergence properties. On the other hand, in the nonlinear filtering systems such as in SLAM, the EKF is a common tool. EKF introduces a step of linearization for the nonlinear systems, and a first-order Taylor expansion performs linearization around the current estimate. The optimality of EKF's is shown as long as linearization is performed around the state vector's exact value. It is the value to estimate in practice and is therefore not usable, and this can lead to problems of accuracy. Nonetheless, estimates are close enough to the reality, for the most part, to allow the EKF to be used.

KF is Bayes filters which signify posteriors by using the Gaussians [16], for example, the distributions of unimodal multivariate that can be denoted efficiently by a minor sum of parameters. The KF SLAM is based on the hypothesis that the transformation and estimation functions are linear with the introduction of Gaussian noise. In state-of-the-art SLAM, KF has two main variations. The first one is the EKF, and the second one is the information filtering (IF) or EIF. The EKF is usually applicable for the nonlinear functions by approximating the mobile robot motion model by means of linear functions. A variety of the SLAM algorithms use the EKF and IF applied by propagating the state error covariance inverse [17–19]. IF is more advantageous as compared to the KF. Initially, the information is filtered out by summing the vector and matrices of information which resultantly give a more precise estimate. Next, the IF is steadier than the KF. Lastly, the EKF is comparatively slow while estimating the maps of having dimensions, because the measurement of every vehicle normally affects the Gaussian parameters. Consequently, the updates need prohibitive times when faced with a situation having several landmarks.

In recent years, the SLAM and autonomous mobile robot combinations play an important role in the controlling disaster field. Particularly, the autonomous robots are widely used for the maintenance and rescue operations in the disaster controlling such as radioactivity leaks. Since the area is unreachable, simultaneous mapping of the environment and the robot localization is crucial to determine the exact source spot [20–23]. Therefore, SLAM has been an important issue as the localization degree hangs on active mapping. However, the SLAM implementation by using the EKF is pretty exciting because of the approximation of the sensor noises and real-time stochastic system as Gaussian. Therefore, inappropriate alteration of the noise covariance may result in filter divergence over time, resulting in the complete system becoming unstable. The researchers presented some alternate methods that are moderately straightforward but severe computationally which have the benefit to accommodate the noise model other than the Gaussian such as UKF, FastSLAM, and Monte Carlo localization [24–26].

1.1. Contributions. In the above paragraphs, the authors investigated the SLAM with KF and EKF. The performance of such models under localization is not yet well-thought-out. Therefore, in this work, the authors analyzed SLAM by suing linear KF and EKF. The basic contribution of this work included one dimensional (1D) SLAM using a linear KF (a) motionless robot with absolute measurement, (b) moving vehicle with absolute measurement, (c) motionless robot with relative measurement, (d) moving vehicle with relative measurement, and (e) moving vehicle with relative measurement while the robot location is not detected. Furthermore, the authors analyzed the localization performance of SLAM with EKF. The proposed SLAM-based algorithms are evaluated and compared with each other and also with other algorithms regarding SLAM. More precisely, the proposed SLAM algorithms present good accuracy while maintaining a sensible computational complication.

1.2. Organization of the Paper. The structure of this paper is as follows: Section 2 demonstrates the work related to SLAM and Section 3 demonstrates the proposed SLAM algorithms. The subsections of Section 3 are SLAM with KF and SLAM with EKF, respectively. Section 4 demonstrates the comparison of the proposed and other algorithms. Finally, Section 5 demonstrates the conclusion and future direction of the proposed algorithms.

2. Related Work

Before presenting the proposed SLAM algorithms, it would be better to present some background knowledge and related work on SLAM algorithms. In this section, the authors present a detailed description of the SLAM that forms the basis of the proposed SLAM algorithms. Compared to the current solutions, many people still do not have highly accurate instruments; they still have challenging piloting capabilities and can solve the SLAM problem. Their mapping, therefore, depends on the toughness policy of acting as a replacement for the accurate world definition. With linear KF, this approach is a new research concept for SLAM.

Regarding the SLAM, readers may not be familiar with the origin and its derivation may refer to the standard and current work on SLAM [27, 28]. For the SLAM problem, the first method was introduced between 1986 and 1991. Smith and Chesseman [29] published a paper in 1986 for the solution of SLAM problems. They present the EKF to solve this problem. In that paper, they established a numerical basis for explaining the relation between landmarks and operating the geometric uncertainty.

Several other researchers have worked on various SLAM issues. For example, in [30–32], the authors presented a new architecture that applies one monocular SLAM system for the tracking of unconstraint motion of the mobile robot. The improved oriented FAST and rotated BRIEF (ORB) characteristics show the landmarks to design a network feature procedure of detection. An enhanced matching feature system has enhanced function matching strength. The updated EKF measures the free-moving visual sensor's multiple dimensional states rather than the standard EKF. Furthermore, in [33, 34], the authors address the issue of the applications of SLAM for navigation problems. For the solution of high-accuracy problems, an EKF or particle filter (PF) algorithm [35] is frequently applied to the processing of data. The PF algorithm, which is often applied for the G-mapping SLAM technique, is well-matched for the nonlinear system's investigation. Though, PF computational dimensions are larger than those of EKF. Therefore, EKF and PF also have some disadvantages in the process of navigation. The Gaussian smoothing filter and its modification are used which is based on the distributed computing scheme. This algorithm is meaningfully better to the EKF and PF algorithms regarding the computational speed.

For the reduction of the linearization error of KF algorithms, the authors presented three techniques and their viability and efficiency are assessed by SLAM [36]. In the derivative-based approaches of the KF system, the linearization error is undetectable owing to the practice of the Taylor expansion for the linearization of the nonlinear motion process. The presented three techniques reduce the error of linearization by substituting the Jacobian observation matrix with new formulations. Similarly, in [37], a SLAM with limited sensing by applying EKF is proposed. The robot's problem with creating a map of an unidentified atmosphere while adjusting its particular location which is the basis on a similar map and sensor information is called SLAM. Because sensor accuracy plays a major part in this issue, most of the planned schemes comprise the use of high-priced laser sensor nodes and comparatively innovative and inexpensive RGB-D cameras. These sensors are too costly for some applications, and RGB-D cameras consume much power, CPU, or communication specifications for on-board or PC processing of data. Thus, the authors tried to model an uncertain setting using a low-cost device, EKF, and dimensional features such as walls and furniture.

Usually, the typical filter uses the scheme model and former stochastic info to approximate the subsequent robot state. Though in the real-time condition, the sound statistics possessions are comparatively unidentified, and the system is imprecisely demonstrated. Therefore, the filter deviation might arise in the incorporation scheme. Furthermore, the predictable precision might be stimulating to be grasped due to the nonappearance of the receptive time-varying of mutually the process and measurement noise statistic. So, the outdated approach desires to be upgraded pointing to deliver an aptitude to guesstimate those belongings. To solve this problem, the new adaptive filter is proposed in [38] named as an adaptive smooth variable structure filter (ASVSF). The upgraded SVSF is consequential and executed; the process and measurement noise statistics are appraised by using the maximum a posteriori creation and the weighted exponent concept. The authors applied ASVSF to overwhelm the SLAM problem of a self-directed mobile robot; hereafter, it is shortened as an ASVSF-SLAM algorithm.

In [39], the authors presented a 3D cooperative SLAM for a joint air grounded robotic system which is intended to succeed an indoor quadrotor flying done composed with a Mecanum-wheeled omnidirectional robot (MWOR) in indoor unidentified and no GPS environments. Moreover, an Oriented Fast and Rotated BRIEF- (ORB-) SLAM 2.0 method is applied to yield a 3D chart and determine concurrently the location of the indoor quadrotor, and a particle-filter SLAM (FastSLAM 2.0) method is applied to shape the 2D chart of the universal atmosphere for the MWOR. An additional accurate 3D quadrotor location estimation technique for the quadrotor is planned with the help of the MWOR. A cooperative SLAM applying fuzzy Kalman filtering is presented to fuse the productions of the ORB-SLAM 2.0, FastSLAM 2.0, and quadrotor location estimation methods, in order to localize the quadrotor further precisely. Mutually, SLAM methods, quadrotor position estimation method, and cooperative SLAM have been executed in the robotic operation system atmosphere.

Recent work on SLAM [40] attempted to address the issue of SLAM landmarks [41]. The authors presented an AUV vision-based SLAM, in which the submerged nonnatural landmarks are utilized for visual sensing of onward and down cameras. The camera can also estimate the AUV location data, along with several navigation sensor nodes such as depth sensor node, Doppler velocity log (DVL), and an inertial measurement unit (IMU). The landmark detection algorithm is organized in a framework of conventional EKF SLAM to measure the landmark and robot status. Furthermore, partial observability of mobile robot based on EKF is explored in [42, 43] to find the answer that can avoid erroneous measurements. When considering only certain

environmental landmarks, the computational costs of mobile robots can be minimized, but with an increase in device uncertainties. The fuzzy logic methodology is presented to guarantee that the calculation has attained the desired output even though some of the landmarks have been omitted for reference purposes. For the measurement invention of KF, fuzzy logic is used to exact the location of the mobile robots and any sensed landmarks all throughout the process of observations.

Researchers have proposed several algorithms for SLAM; some of which are already discussed in the above pages. Most of them focused on the landmark's estimation, performance, accuracy, and effectiveness of the SLAM algorithm. However, there are still some important and fundamental issues that need to be addressed, such as an optimal solution for SLAM, active SLAM for SLAM development, SLAM failure detection, SLAM front end robust algorithm, and SLAM algorithm that considers various aspects at once. Therefore, in this paper, the authors attempted to propose a modified SLAM algorithm by applying KF and EKF. The authors presented SLAM algorithms that consider several aspects of the SLAM such as velocity, distance, coverage area, maximum range, and localization time. The notations used in this work are listed in Table 1.

Table 1: Index of notation.

Notation	Description
X_k	Current state
X_{k-1}	Previous state
Q_k	Covariance matrix of prediction
R_k	Covariance matrix of observation
X_{k+1}	Estimated state vector
$P_{k+1/k}$	Covariance matrix for prediction
Q	Process noise matrix
R	Measurement noise matrix
$w_k \sim N(0, Q_k)$	Process noise
$v_k \sim N(0, R_k)$	Observation noise
$K+1$	Time instant
Z_{k+1}	Estimated measurement vector
$\nabla F_x, \nabla F_u$	Jacobian matrices of the function f
B, F	State transition matrix
U	Input of the process noise
W, V	Vectors of state measurement noise
v	Mobile robot velocity
p	Landmark position
JF	Jacobian of state equation
K_{k+1}	Kalman gain
$X_{k+1/k+1}$	Updated estimate
$P_{k+1/k+1}$	Updated covariance
H	Measurement Jacobian or linearization matrix
dt	Global time
T	Time
t	Initial time
LM	Landmark
P_k	New state covariance matrix

3. Proposed Simultaneous Localization and Mapping Algorithms

This section presents the proposed SLAM algorithms based on KF and EKF. The proposed algorithms are analyzed and evaluated in the next subsections.

3.1. Simultaneous Localization and Mapping with Kalman Filter. In the following section, the authors presented the theory of SLAM which results in efficient localization and mapping in WSNs. Specifically, the author presents the analysis of the operating environment and finally discussed the proposed algorithm and compared it with other SLAM algorithms. In the existence of Gaussian white noise, the KF provides a well-designed and statically optimum explanation for the linear systems. It is a technique that uses linear estimation associated with the states and error covariance matrixes for the purpose to produce gain stated to as the Kalman gain. Such benefit is added to the estimation of a preceding condition, thereby generating an estimate of a posteriori [44]. The below equations define the dynamic model of the system and the measuring model used for the linear state approximation in general which consists of two f and h functions.

$$Xk+1 = f(Xk, Uk, Wk), \tag{1}$$

$$Zk+1 = h(Xk, Vk), \tag{2}$$

which administrate state proliferation and state measurements, where U is the input of the process, W and V are the vectors of state and measurement noise, while k represents the discrete-time. In the above equations, f and h are typically based on a set of discretized difference equations that govern the dynamics and observation from the method.

$$X_{k+1/k} = F_k \times X_{k/k}, \tag{3}$$

$$P_{k+1/k} = F_k \times P_{k/k} \times F_k^T + A_k \times Q_k \times A_k^T, \tag{4}$$

$$K_{k+1} = P_{k+1/k} \times H_k^T \left[H_{k+1} \times P_{k+1/k} \times H_{k+1}^T + R_k\right]^{-1}, \tag{5}$$

$$X_{k+1/k+1} = X_{k+1/k} + K_{k+1}[Z_{k+1} - H_{k+1} \times X_{k+1/k}], \tag{6}$$

$$P_{k+1/k+1} = [I - K_{k+1} \times H_{k+1}] \times P_{k+1/k}. \tag{7}$$

Such equations from the KF-based method are used iteratively in conjunction with Equations (1) and (2). Equation (3) generalizes the prior state estimate, and Equation (4) represents the equivalent state covariance error. The gain of Kalman can be estimated by Equation (5) which is applied to update the state approximation and covariance

error, defined by Equations (6) and (7), correspondingly. EKF is practically comparable to the iterative KF method, and sometimes, it is used for the nonlinear systems. By applying the Jacobian, which is a first-order partial derivative, the measurement H_{k+1} and nonlinear system F_k matrices are linearized.

A one-dimensional SLAM with KF is applied for a motionless robot, and the measurement is considered an absolute measurement. A 1-DoF mobile robot is used which is motionless in a fixed position of a straight line. The mobile robot is used for detecting the motionless/stationary landmarks. The mobile robot velocity v and position p of the landmarks are calculated by applying SLAM with linear KF. Ten numbers of landmark positions are considered. For the real trajectory, the robot is motionless at a given position which is $v = 0$ m/s. The landmark distance is relative to the mobile robot's location/position which had a moderate measurement noise as shown in Figure 1. The state vector is the diagonal of those that correspond to the robot's present state by projecting the next one. The vector used for the control is null; it shows that there are no exterior inputs to vary the mobile robot's state; i.e, the velocity and position of the robot are constant. The initial matrix of covariance is not prevalent; it is characterized by a broad diagonal ambiguity in both the robot's landmark location and state and equal ambiguity/uncertainty. The result of mobile robot localization with absolute measurement is shown in Figure 2.

Next, a one-dimensional SLAM with KF is applied for a moving vehicle and the measurement is considered an absolute measurement. A 1-DoF mobile robot is traveling on a straight path. The mobile robot is sensing the motionless/stationary landmarks. The robot velocity v and the landmark position/velocity p are calculated by applying SLAM using a linear KF, and in this case, all the measurements are absolute, see Figure 3. The landmark positions are the same as the previous one.

Furthermore, a one-dimensional SLAM with KF is applied for a motionless robot, and the measurement is considered a relative measurement. Here, a 1-DoF mobile robot is used in a motionless and fixed position of a straight lane that detects the motionless/stationary landmarks. The robot velocity and the position/location landmarks are calculated by using the SLAM with a KF, see Figure 4. The fourth one is a one-dimensional SLAM with linear KF. In this case, a moving vehicle is considered with a relative measurement and a 1-DoF robot is traveling on a straight line that detects the motionless/stationary landmarks. The robot position/location, velocity, and landmark position are calculated through SLAM with linear KF. Here, all the measures are comparative to the position/location of the mobile robot, see Figure 5.

The last one is almost different from the previous four SLAM algorithms. In this case, a one-dimensional SLAM with linear KF is considered and the vehicle is moving with a relative/comparative motion. The position/location of the mobile robot is not observed in this case. A mobile robot is traveling on a straight line that detects the landmarks which are motionless as shown in Figure 6. The velocity of the robot and its landmark are calculated by applying SLAM with

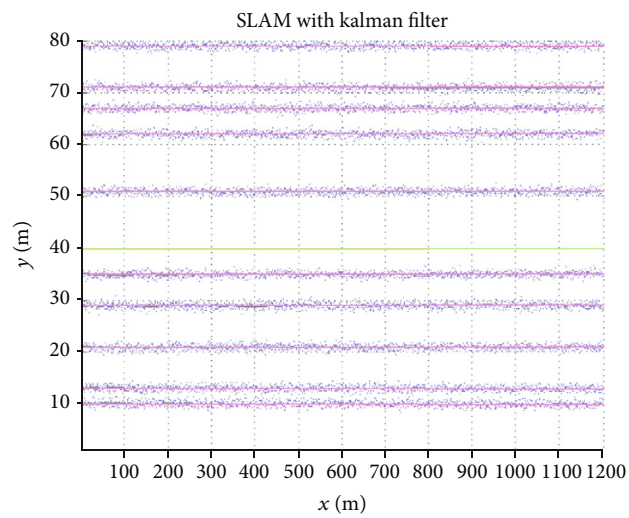


Figure 1: SLAM with motionless robot and absolute measurement while having a moderate measurement noise.

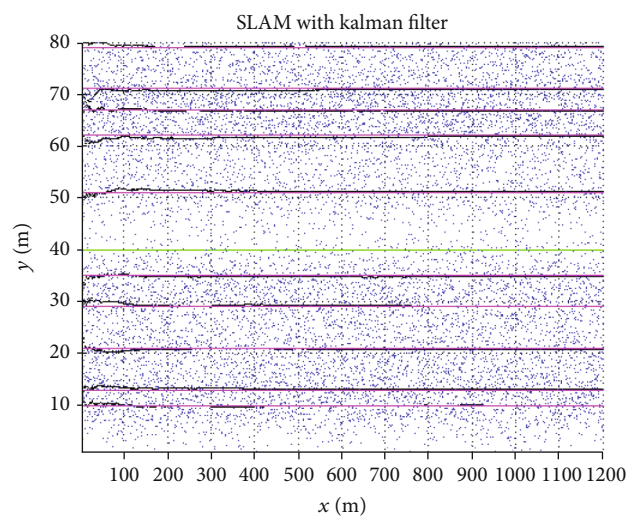


Figure 2: SLAM with motionless robot and absolute measurement.

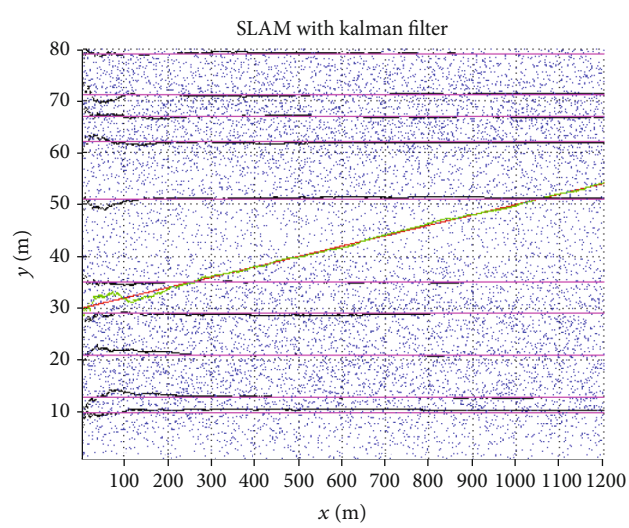


Figure 3: SLAM with moving vehicle and absolute measurement.

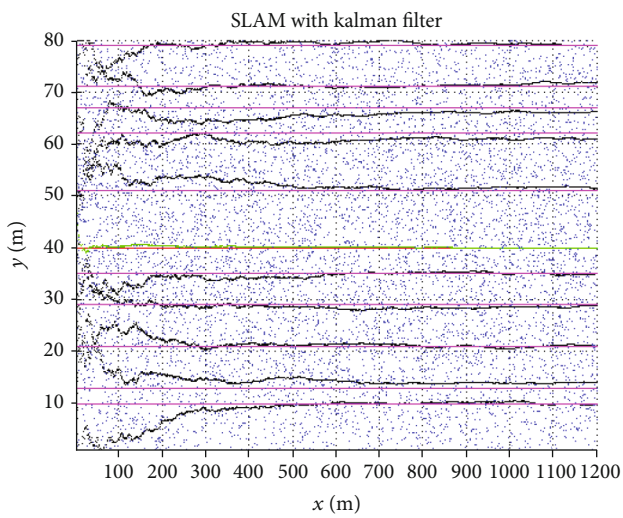


Figure 4: SLAM with motionless robot and relative measurement.

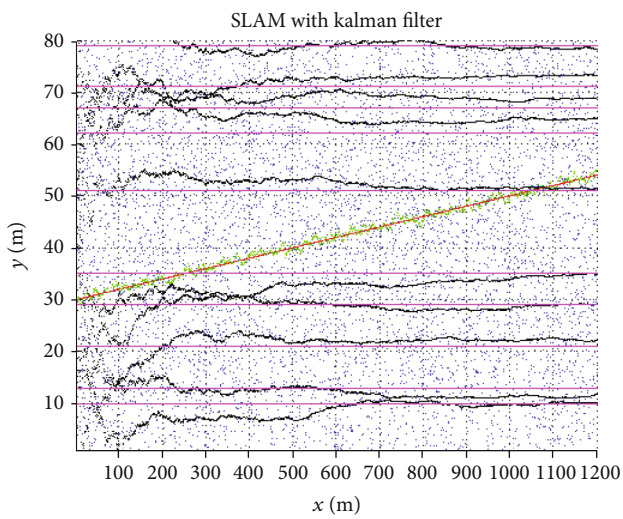


Figure 5: SLAM with moving vehicle and relative measurement.

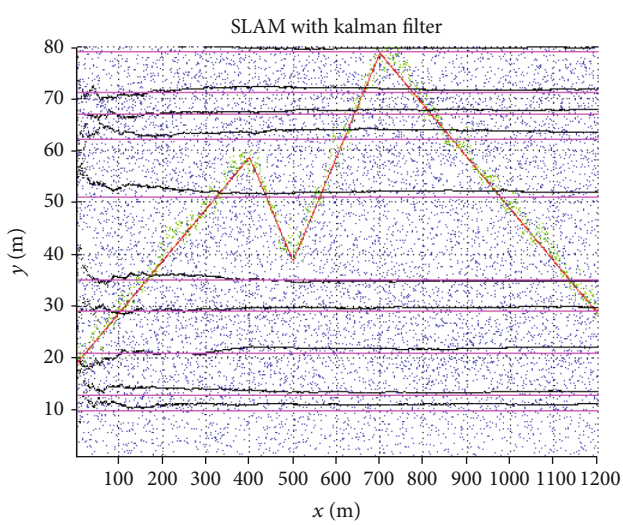


Figure 6: SLAM with moving vehicle and relative measurement while the position of the robot is not observed.

linear KF. As mentioned before, the position is not observed and all the measurements are relative/comparative to the mobile robot position/location.

4. Simulation Results and Discussion

The proposed SLAM algorithm is evaluated by simulation. In this simulation, the author evaluates the SLAM algorithm by conducting a different experiment with different landmarks. The simulation is divided into five steps, such as a motionless robot with absolute measurement, a moving vehicle with absolute measurement, a motionless robot with relative measurement, a moving vehicle with relative measurement, and a moving vehicle with relative measurement while the robot location is not detected. The number of time-stamps is 1200 with the map of dimension [180]. The landmark positions are set to be LM = 10 which are denoted by p. For the real trajectory, the velocity and position are $v = 1.0$ m/s and $p = 40$, respectively, at state $x(k) = x(k-1)$ and $v(k) = v(k-1)$, i.e., motionless at a given position having a moderate measurement noise as shown in Figure 1. The process p_n and measurement noise r is added, and the landmark distance is relative to the robot position, see Figure 2. The state equation is a diagonal of those, which ensures that the next state's estimate or prediction is equal to the present state. The control vector is null; it shows that there are no exterior inputs that vary the state of the robot because, as we stated earlier, the velocity v and position p are constant. Also, the primary covariance matrix is well-defined by a higher diagonal uncertainty mutually in the position of the landmark and the robot state and by a comparable uncertainty, which means that none prevails over the other.

The process noise matrix represented by Q and the measurement noise matrix represented by R are computed in which the landmarks are motionless. For the next state prediction, the measurement is done at the prediction position, and for observation, it is measured at the right position/location $x(k+1)$, $v(\mathrm{k}+1)$, and p. The landmark positions are similar for all five methods. However, for this case, a vehicle is considered with constant velocity $v = 0.02$ m/s and the position are $p = 30$. Also, in this case, the landmark distance is absolute. In the third case, the robot is motionless and the measurement is relative at a given velocity and position $v = 1.0$ m/s and $p = 40$, respectively. The fourth one is the SLAM with linear KF in which the vehicle is moving and the measurement is relative. The landmark distance is relative to robot position and a vehicle with a constant velocity of $v = 0.02$ m/s and at the position, see Figure 5, the red line denotes the position. The last one is the SLAM with linear KF and a vehicle is moving, and the measurement is relative. Note, in this case, the position is not observed as the previous. The constant velocity of the vehicle is set to be $v = -0.2$ m/s and the position is 20, as can be seen in Figure 6. The above-mentioned algorithms for SLAM with KF are evaluated in deep detail. The authors considered a variety of aspects regarding the SLAM localization. The offered SLAM algorithms present a high level of accuracy in various conditions and perform well in terms of velocity, distance, coverage area, etc.

4.1. Simultaneous Localization and Mapping with Extended Kalman Filter. EKF is well-known as a widespread resolution to the SLAM problem for mobile robot localization. In this section, the authors realized the EKF SLAM-based algorithm for a mobile robot that follows a specific trajectory. EKF SLAM relies on present elements of the navigation system known as landmarks to change the location of the robot. EKF SLAM for a mobile robot is executed in a defined field with a specific feature. The authors considered two basic mathematical models such as the EKF state and observation model that are represented below.

$$X_{k+1} = f(X_k, U_k, w_k), \tag{8}$$

$$Z_{k+1} = h(X_{k+1}, v_{k+1}), \tag{9}$$

where $w_k \sim N(0, O_k)$ and $vk \sim N(0, Rk)$ which characterize the process and observation noise. Here, X_{k+1} denotes the estimated state vector at time $k+1$. The time is the discrete time for a known input U_k assuming all noise to be w_k. In Equation (9), Z_{k+1} represents the estimated measuring vector at the time instant $k+1$, where v_k is the observation noise. Q_k and R_k denote the covariance matrix of prediction and observation, respectively. EKF offers an approximation of the optimal state estimate. The EKF-SLAM objectives are to estimate recursively the landmark state X_k as stated by the Z_{k+1} measurement. EKF is basically divided into several steps which are represented as at the initial state, the state vector X_{k+1} will become

$$X_{k+1} = f(X_k, U_{k+1}) + \nabla F_x \times \left(X_k - X_{\frac{k}{\hat{k}}}\right). \tag{10}$$

In the prediction stage, the covariance matrix for prediction $P_{k+1/k}$ can be represented as

$$P_{k+1/k} = \nabla F_x \times P_{k/k} \times \nabla F_x^T + \nabla F_u \times Q_k \times \nabla F_u^T. \tag{11}$$

In the above equation, the ∇F_x and ∇F_u denote the Jacobian matrices of the function f concerning the state vector which is X_{k+1}. The state transition matrix is denoted by B, and F is the state equation which can be represented as follows:

$$B = \begin{bmatrix} dt \times \cos(x(3)) & 0 \\ dt \times \cos(x(3)) & 0 \\ 0 & dt \end{bmatrix} \tag{12}$$

$$F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{13}$$

Therefore, the Jacobian of the state equation will become

$$JF = \begin{bmatrix} 0 & 0 & -dt \times u(1) \times \sin(x(3)) \\ 0 & 0 & dt \times u(1) \times \cos(x(3)) \\ 0 & 0 & 0 \end{bmatrix}, \tag{14}$$

and the global initialization Jacobian G can be written as follows:

$$G = \begin{bmatrix} -\sqrt{\delta(1)} & -\sqrt{\delta(2)} & 0 & \sqrt{\delta(1)} & \sqrt{\delta(2)} \\ \delta(2) & -\delta(1) & -1 & -\delta(2) & \delta(1) \end{bmatrix} \tag{15}$$

In the observation and update phase, the observation model Z_{k+1} at $X_{k+1/k}$ can be represented as

$$Z_{k+1} = h(X_{k+1/k}) + H_{k+1} \times (X_{k+1/k} - X_{k+1}). \tag{16}$$

To apply the KF update cycle, i.e., $X_{k+1/k}$ and $P_{k+1/k}$, the KF gain can be computed.

$$K_{k+1} = P_{k+1/k} \times H_{k+1}^T \times \left[H_{k+1} \times P_{k+1/k} \times H_{k+1}^T + R_k\right]^{-1}, \tag{17}$$

where K_{k+1} is the Kalman gain. With measurement of Z_{k+1}, the updated estimate can be

$$K_{k+1/k+1} = X_{k+1/k} + K_{k+1} \times [Z_{k+1} - H_{k+1} \times X_{k+1/k}]. \tag{18}$$

If the Z_k of measurement is available, EKF calculates the matrix of Kalman gain and integrates the invention of measurement to obtain the approximate state X_k, accompanied by the update of the state error matrix. Therefore, the measurement updated step from the above equation will become

$$X_k = X_k + K_k \times [Z_k - h(X_k)], \tag{19}$$

$$P_k = [I - K_k \times H_k] \times P_k. \tag{20}$$

Therefore, the update covariance $P_{k+1/k+1}$ can be represented as

$$P_{k+1/k+1} = [I - K_{k+1} \times H_{k+1}] \times P_{k+1/k}, \tag{21}$$

where K_k is the Kalman gain and P_k is the new state covariance matrix. H is the measurement Jacobian or linearization matrix and X_k denotes the state vector X_k estimate.

After evaluating EKF in deep detail, the authors conclude that the EKF also has some disadvantages that is if the process and measurement noise are not accurately displayed, the robot will diverge from its route which resultantly give a contradiction. Particularly, in the case of the robot velocity, the robot is sensitive to the velocity as by varying the velocity the robot is diverging from its route as shown in Figure 7 . However, in the first case, the velocity is $v = 1.0$ m/s as shown in Figure 8.

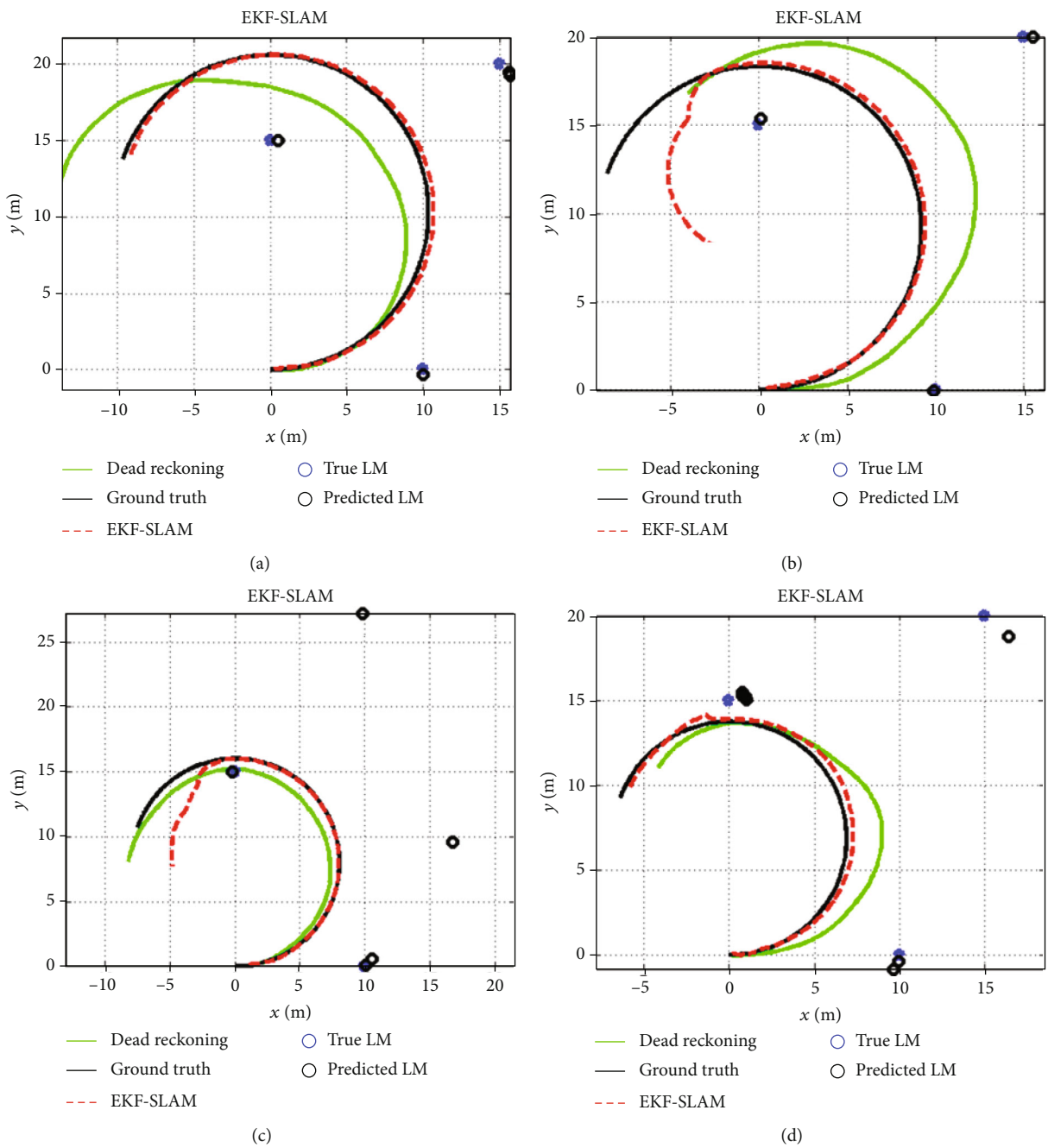


Figure 7: The SLAM EKF performance for various v values: (a) $v = 0.9$ m/s, (b) ˙, (c) $v = 0.7$ m/s, and (d) $v = 0.6$ m/s, where the green line represents the dead reckoning calculation of the existing location, which is the outcome of the displacement function applied for the previous location. The black mark is the ground truth information that is necessary for real-time localization. The red mark is the EKF-SLAM. The blue asterisks represent the true landmarks, and the black circle is the estimated landmarks.

5. Simulation Results and Discussion

The proposed SLAM EKF algorithm is evaluated through simulation. In this simulation, the author evaluates the SLAM EKF algorithm by performing simulation with various factors. Firstly, the time is $t = 0$, end time is $t = 60$ sec, while the global time is $\mathrm{dt} = 0.1$ sec. In this simulation, the state vector is considered $x\mathrm{Est} = [0\,0\,0]^T$ in which the $x\mathrm{True} = x$ Est, while at the dead reckoning state $xd = x\mathrm{True}$. The landmark coordinates are [xy], i.e., $\mathrm{LM} = [0\,15\,;10\,0\,;15\,20]$. The maximum range is set to be 20 at the initial stage and parameter $\alpha = 1$. For the input parameters, the time is set to be $T = 10$ sec, the velocity is $v = 1.0$ m/s, and yaw rate = 5 deg/sec. At the initial stage, the velocity is limited to $v = 1.0$ m/s as can be seen in Figure 8; however, in the next stage, the velocity is varying.

In the case of varying the velocities as can be seen in Figure 7, the velocities are set to be $v = 0.9$ m/s, $v = 0.8$ m/s,

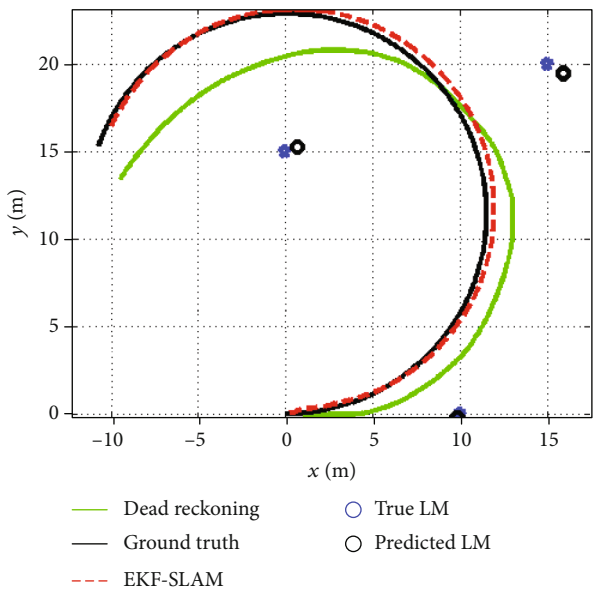


Figure 8: Performance of SLAM with Extended Kalman Filter.

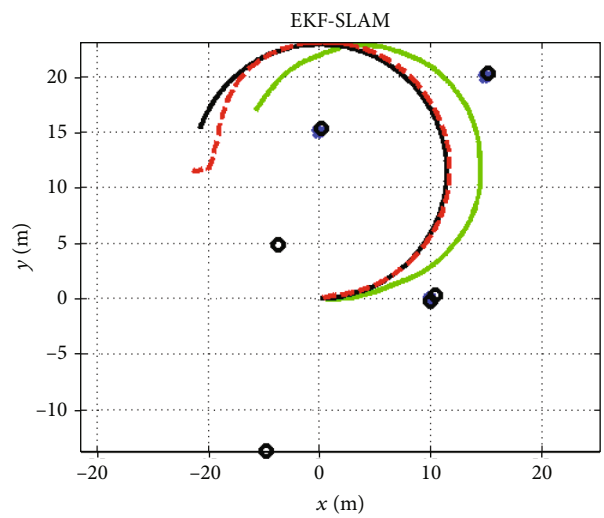


Figure 9: Performance of SLAM with Extended Kalman Filter in case of higher range.

$v = 0.7$ m/s, and $v = 0.6$ m/s. By varying the velocity of the robot, the robot is diverging from its route and, therefore, reduces the coverage area as can be seen in Figure 7(a)-7(d). Also, the error between the true landmark and predicted landmark is increasing. On the other hand, for higher velocities (more than $v = 1.0$ m/s), the proposed algorithm is not applicable, because in the SLAM, the robot is following the prior defined map and the robot keeps communication with the surrounding. However, in our previous study, we mentioned the higher velocities for the robot, in the case of EKF, UKF, and PF, the coverage area, and localization were increasing by increasing the velocity. Furthermore, the maximum range was set to be 20 as shown in Figure 6, but by modifying the maximum range to 30 or above, in this case also, the robot diverges from its route of localization as shown in Figure 9. Resultantly, the authors conclude that the proposed algorithm is more suitable for constant velocity which presents a high level of accuracy.

6. Comparison of the Proposed and Other Algorithms

In the above sections, the authors investigated and evaluated well about the proposed SLAM algorithms. However, to demonstrate the effectiveness and better performance of the planned algorithms, the authors present a brief comparison of the proposed algorithms with other algorithms in this section.

In [45], the authors presented a neurofuzzy-based adaptive EKF method. The purpose of this method is to estimate the right value of matrix R at every stage. They plan an adaptive neurofuzzy EKF to lessen the variance among the theoretical and actual covariance matrices. The parameters for this technique are then skilled offline by using a particle swarm optimization method. A mobile robot steering with a number of landmarks under two situations is assessed. The technique is applied that the adaptive neurofuzzy EKF provides development in performance effectiveness.

An EKF-based SLAM system for a mobile robot with sensor bias estimation is presented in [46]. The authors proposed an improved method for EKF which is practical to the issue of mobile robot SLAM which has taken into consideration the sensor bias issue. Mobile robot Pioneer 3-AT is taken as the model for studying the theoretical derivation and the authentication of the investigation in this work. At first, the kinematic model of Pioneer 3-AT mobile robot is introduced; then, the improved EKF method, taking into account the issue of bias estimation and compensation, is anticipated to increase the precision of the location estimation. In addition, a study explores the autonomous location and atmosphere mapping of stirring substances under the dust and low lighting situations in underground underpasses. The typical EKF algorithm has a problem that machine noise and the prior statistical characteristics of the observed noise cannot be predicted accurately. Thus, the authors presented an enhanced EKF algorithm to accomplish a fuzzy adaptive SLAM [45, 47, 48]. Therefore, to predict the position, a laser matching is applied to the EKF prediction process, and the weighted average location is used as the final location of the predicted component. The machine noise and the weighted value of experiential noise become fuzzily recognizable by observing the variation of mean value and covariance. The improved filtering algorithm is applied to a SLAM simulation study and measure the impact on position estimation of four dissimilar landmark measurements.

A recent approach strong tracking second-order (STSO) central difference SLAM is presented in [49] which it is based on the tracking second-order central difference KF. The proposed procedure gathers the second-order central differential filter (SOCDF), strong tracking filter (STF), and PF. Using Cholesky decomposition, the algorithm uses the Sterling Interpolation second-order method to solve a nonlinear system problem. This methodology transmits directly in the probabilistic estimation of SLAM by adding the covariance

square root factor. Furthermore, in [50], a visual-inertial SLAM feedback mechanism is presented for the real-time motion assessment of the SLAM map. The main aspect of this mechanism is that the front-end and the back-end can support each other in the VISLAM. The output from the back-end is fed to the KF-based front-end to decrease the motion estimation error produced by the linearization of the KF estimator. On the other hand, this more accurate front-end motion estimation will improve back-end optimization as it provides the back-end with an exact primary state. Similarly, the EKF-based SLAM approaches are presented in [33, 51, 52] which focus on the performance and effectiveness of the SLAM. Each algorithm presents well in its domain, but the proposed SLAM algorithms perform well compared to the other SLAM algorithms.

7. Conclusion and Future Directions

In this work, the SLAM algorithm is proposed in two different methods such as SLAM with linear KF and SLAM with EKF. Firstly, SLAM with linear KF is implemented in five different methods such as the motionless robot with absolute measurement, moving vehicle with absolute measurement, a motionless robot with relative measurement, moving vehicle with relative measurement, and moving vehicle with relative measurement while the robot location is not detected. The landmark position was set to be 10 for all five cases. The mobile robot position or velocity and landmark position are calculated by applying SLAM using a linear KF. Secondly, the SLAM with EKF is implemented and an analytical expression for the EKF-based SLAM algorithm is derived and their presentation is evaluated. The SLAM algorithm with EKF is evaluated in various scenarios, and several iterations are applied to explain the performance of EKF-based SLAM well. The proposed algorithm is simulated for varying velocities, and their performance is presented in Figure 8. Each process of localization is effective in its domain. In this analysis, many localization factors such as velocity, coverage area, localization time, and cross section area are taken into consideration. The proposed SLAM-based algorithm performance is intensively assessed by executing numerous iterations as can be seen in the figures above. The planned SLAM-based algorithms present a high precision while preserving realistic computational complexity. The simulation outcomes indicate that the planned SLAM algorithms can accurately locate the landmark and mobile robot.

Future research will use more simulation and tests to show the robustness of the SLAM in different scenarios and landmarks. We will try to make a robot pilot more originally and also apply SLAM with UKF and PF algorithms. In addition, improving the development of some standards for estimating SLAM approaches, particularly for large-scale SLAM in dynamic situations, is also important to make the results of SLAM algorithms more valuable.

Data Availability

Since the funding project is not closed and related patents have been evaluated, the simulation data used to support the findings of this study are currently under embargo while the research findings are commercialized. Requests for data, based on the approval of patents after project closure, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is supported by the National Key Research and Development Program under Grant 2018YFC0407101 and in part by the National Natural Science Foundation of China under Grant 61801166. It was also supported by the Fundamental Research Funds for the Central Universities under Grant 2019B22214 and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2018R1D1A1B07043331.

References

[1] I. Ullah, Y. Liu, X. Su, and P. Kim, "Efficient and accurate target localization in underwater environment," *IEEE Access*, vol. 7, pp. 101415–101426, 2019.

[2] K. Sha, T. A. Yang, W. Wei, and S. Davari, *A survey of edge computing-based designs for iot security*, Digital Communications and Networks, 2019.

[3] G. Zand, M. Taherkhani, and R. Safabakhsh, "A novel framework for simultaneous localization and mapping," in *2015 Signal Processing and Intelligent Systems Conference (SPIS)*, pp. 109–113, Tehran, Iran, December 2015.

[4] X. Ma, R. Wang, Y. Zhang, C. Jiang, and H. Abbas, "A name disambiguation module for intelligent robotic consultant in industrial internet of things," *Mechanical Systems and Signal Processing*, vol. 136, article 106413, 2020.

[5] T. Pire, T. Fischer, J. Civera, P. De Cristóforis, and J. J. Berlles, "Stereo parallel tracking and mapping for robot localization," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1373–1378, Hamburg, Germany, October 2015.

[6] T. A. Johansen and E. Brekke, "Globally exponentially stable Kalman filtering for slam with ahrs," in *2016 19th International Conference on Information Fusion (FUSION)*, pp. 909–916, Heidelberg, Germany, July 2016.

[7] S. Safavat, N. N. Sapavath, and D. B. Rawat, "Recent advances in mobile edge computing and content caching," *Digital Communications and Networks*, 2019.

[8] S. Fu, H.-y. Liu, L.-f. Gao, and Y.-x. Gai, "Slam for mobile robots using laser range finder and monocular vision," in *2007 14th International Conference on Mechatronics and Machine Vision in Practice*, pp. 91–96, Xiamen, China, December 2007.

[9] C. Kerl, J. Sturm, and D. Cremers, "Dense visual slam for rgb-d cameras," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2100–2106, Tokyo, Japan, November 2013.

[10] Y. Li, S. Xia, M. Zheng, B. Cao, and Q. Liu, "Lyapunov optimization based trade-off policy for mobile cloud offloading in

heterogeneous wireless networks," *IEEE Transactions on Cloud Computing*, 2019.

[11] C. Cadena, L. Carlone, H. Carrillo et al., "Past, present, and future of simultaneous localization and mapping: toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[12] M. Raja, "Application of cognitive radio and interference cancellation in the l-band based on future air-to-ground communication systems," *Digital Communications and Networks*, vol. 5, no. 2, pp. 111–120, 2019.

[13] I. Ullah, J. Chen, X. Su, C. Esposito, and C. Choi, "Localization and detection of targets in underwater wireless sensor using distance and angle based algorithms," *IEEE Access*, vol. 7, pp. 45693–45704, 2019.

[14] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: a survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[15] Y. Zhang, H. Wen, F. Qiu, Z. Wang, and H. Abbas, "Ibike: intelligent public bicycle services assisted by data analytics," *Future Generation Computer Systems*, vol. 95, pp. 187–197, 2019.

[16] J. Aulinas, Y. R. Petillot, J. Salvi, and X. Lladó, "The slam problem: a survey," *CCIA*, vol. 184, no. 1, pp. 363–371, 2008.

[17] P. Jensfelt, D. Kragic, J. Folkesson, and M. Bjorkman, "A framework for vision based bearing only 3d slam," in *Proceedings 2006 IEEE International Conference on Robotics and Automation 2006. ICRA 2006*, pp. 1944–1950, Orlando, FL, USA, May 2006.

[18] A. J. Davison and D. W. Murray, "Simultaneous localization and map-building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 865–880, 2002.

[19] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *The international Journal of robotics Research*, vol. 21, no. 8, pp. 735–758, 2016.

[20] M. N. Santhanakrishnan, J. B. B. Rayappan, and R. Kannan, "Implementation of extended kalman filter-based simultaneous localization and mapping: a point feature approach," *Sādhanā*, vol. 42, no. 9, pp. 1495–1504, 2017.

[21] T. Rahman, X. Yao, and G. Tao, "Consistent data collection and assortment in the progression of continuous objects in iot," *IEEE Access*, vol. 6, pp. 51875–51885, 2018.

[22] Y. Li, J. Liu, B. Cao, and C. Wang, "Joint optimization of radio and virtual machine resources with uncertain user demands in mobile cloud computing," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2427–2438, 2018.

[23] X. Su, I. Ullah, X. Liu, and D. Choi, "A review of underwater localization techniques, algorithms, and challenges," *Journal of Sensors*, vol. 2020, 24 pages, 2020.

[24] P. Yang and W. Wu, "Efficient particle filter localization algorithm in dense passive rfid tag environment," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 10, pp. 5641–5651, 2014.

[25] P. Yang, "Efficient particle filter algorithm for ultrasonic sensor-based 2d range-only simultaneous localisation and mapping application," *IET Wireless Sensor Systems*, vol. 2, no. 4, pp. 394–401, 2012.

[26] G. Cotugno, L. D'Alfonso, W. Lucia, P. Muraca, and P. Pugliese, "Extended and unscented kalman filters for mobile robot localization and environment reconstruction," in *21st Mediterranean Conference on Control and Automation*, pp. 19–26, Chania, Greece, June 2013.

[27] S. Huang and G. Dissanayake, "A critique of current developments in simultaneous localization and mapping," *International Journal of Advanced Robotic Systems*, vol. 13, no. 5, article 1729881416669482, 2016.

[28] G. Dissanayake, S. Huang, Z. Wang, and R. Ranasinghe, "A review of recent developments in simultaneous localization and mapping," in *2011 6th International Conference on Industrial and Information Systems*, pp. 477–482, Kandy, Sri Lanka, August 2011.

[29] R. C. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *The International Journal of Robotics Research*, vol. 5, no. 4, pp. 56–68, 2016.

[30] K.-K. Tseng, J. Li, Y. Chang, K. L. Yung, C. Y. Chan, and C.-Y. Hsu, "A new architecture for simultaneous localization and mapping: an application of a planetary rover," *Enterprise Information Systems*, pp. 1–17, 2019.

[31] Z. Miljković, N. Vuković, and M. Mitić, "Neural extended Kalman filter for monocular slam in indoor environment," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 230, no. 5, pp. 856–866, 2015.

[32] S. Prakash and G. Gu, *Simultaneous localization and mapping with depth prediction using capsule networks for uavs*, 2018, http://arxiv.org/abs/1808.05336.

[33] N. Ayadi, N. Derbel, N. Morette, C. Novales, and G. Poisson, "Simulation and experimental evaluation of the ekf simultaneous localization and mapping algorithm on the wifibot mobile robot," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 8, no. 2, pp. 91–101, 2018.

[34] G. Wang and A. Fomichev, "Simultaneous localization and mapping method for a planet rover based on a gaussian filter," in *InAIP Conference Proceedings*, vol. 2171, no. 1, article 160003, 2019AIP Publishing, 2019.

[35] I. Ullah, Y. Shen, X. Su, C. Esposito, and C. Choi, "A localization based on unscented kalman filter and particle filter localization algorithms," *IEEE Access*, vol. 8, no. 1, pp. 2233–2246, 2020.

[36] F. F. Yadkuri and M. J. Khosrowjerdi, "Methods for improving the linearization problem of extended kalman filter," *Journal of Intelligent & Robotic Systems*, vol. 78, no. 3-4, pp. 485–497, 2015.

[37] O. Ozisik and S. Yavuz, "Simultaneous localization and mapping with limited sensing using extended kalman filter and hough transform," *Tehnicki vjesnik - Technical Gazette*, vol. 23, no. 6, pp. 1731–1738, 2016.

[38] Y. Tian, H. Suwoyo, W. Wang, and L. Li, "An asvsf-slam algorithm with time-varying noise statistics based on map creation and weighted exponent," *Mathematical Problems in Engineering*, vol. 2019, 17 pages, 2019.

[39] C.-C. Tsai, C.-F. Hsu, X.-C. Lin, and F.-C. Tai, "Cooperative slam using fuzzy kalman filtering for a collaborative air-ground robotic system," *Journal of the Chinese Institute of Engineers*, vol. 43, no. 1, pp. 67–79, 2020.

[40] J. Jung, Y. Lee, D. Kim, D. Lee, H. Myung, and H.-T. Choi, "Auv slam using forward/downward looking cameras and artificial landmarks," in *2017 IEEE Underwater Technology (UT)*, pp. 1–3, Busan, South Korea, February 2017.

[41] H. Abdelnasser, R. Mohamed, A. Elgohary et al., "Semantic-slam: using environment landmarks for unsupervised indoor

localization," *IEEE Transactions on Mobile Computing*, vol. 15, no. 7, pp. 1770–1782, 2016.

[42] H. Ahmad, N. A. Othman, and M. S. Ramli, "A solution to partial observability in extended kalman filter mobile robot navigation," *Telkomnika*, vol. 16, no. 1, pp. 134–141, 2018.

[43] P. Thulasiraman and K. A. White, "Topology control of tactical wireless sensor networks using energy efficient zone routing," *Digital Communications and Networks*, vol. 2, no. 1, pp. 1–14, 2016.

[44] F. Demim, A. Nemra, K. Louadj, Z. Mehal, M. Hamerlain, and A. Bazoula, "Simultaneous localization and mapping algorithm for unmanned ground vehicle with svsf filter," in *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*, pp. 155–162, Algiers, Algeria, November 2016.

[45] C. H. Do and H.-Y. Lin, "Incorporating neuro-fuzzy with extended kalman filter for simultaneous localization and mapping," *International Journal of Advanced Robotic Systems*, vol. 16, no. 5, article 1729881419874645, 2019.

[46] X. Xie, Y. Yu, X. Lin, and C. Sun, "An ekf slam algorithm for mobile robot with sensor bias estimation," in *2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 281–285, Hefei, China, May 2017.

[47] Z.-L. Ren, L.-G. Wang, and L. Bi, "Improved extended kalman filter based on fuzzy adaptation for slam in underground tunnels," *International Journal of Precision Engineering and Manufacturing*, vol. 20, no. 12, pp. 2119–2127, 2019.

[48] C. H. Do, H.-Y. Lin, and Y.-C. Huang, "Simultaneous localization and mapping with neuro-fuzzy assisted extended kalman filtering," in *2017 IEEE/SICE International Symposium on System Integration (SII)*, pp. 393–398, Taipei, Taiwan, December 2017.

[49] J. Dai, X. Li, K. Wang, and Y. Liang, "A novel stsoslam algorithm based on strong tracking second order central difference kalman filter," *Robotics and Autonomous Systems*, vol. 116, pp. 114–125, 2019.

[50] J. Bai, J. Gao, Y. Lin, Z. Liu, S. Lian, and D. Liu, "A novel feedback mechanism-based stereo visual-inertial slam," *IEEE Access*, vol. 7, pp. 147721–147731, 2019.

[51] A. Giannitrapani, N. Ceccarelli, F. Scortecci, and A. Garulli, "Comparison of ekf and ukf for spacecraft localization via angle measurements," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, no. 1, pp. 75–84, 2011.

[52] D. Fethi, A. Nemra, K. Louadj, and M. Hamerlain, "Simultaneous localization, mapping, and path planning for unmanned vehicle using optimal control," *Advances in Mechanical Engineering*, vol. 10, no. 1, Article ID 168781401773665, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 3685614, 9 pages
https://doi.org/10.1155/2020/3685614

Research Article

Arabic Sign Language Recognition and Generating Arabic Speech Using Convolutional Neural Network

M. M. Kamruzzaman

Department of Computer and Information Science, Jouf University, Sakaka, Al-Jouf, Saudi Arabia

Correspondence should be addressed to M. M. Kamruzzaman; mmkamruzzaman@ju.edu.sa

Received 28 January 2020; Revised 15 February 2020; Accepted 25 February 2020; Published 23 May 2020

Guest Editor: Yin Zhang

Sign language encompasses the movement of the arms and hands as a means of communication for people with hearing disabilities. An automated sign recognition system requires two main courses of action: the detection of particular features and the categorization of particular input data. In the past, many approaches for classifying and detecting sign languages have been put forward for improving system performance. However, the recent progress in the computer vision field has geared us towards the further exploration of hand signs/gestures' recognition with the aid of deep neural networks. The Arabic sign language has witnessed unprecedented research activities to recognize hand signs and gestures using the deep learning model. A vision-based system by applying CNN for the recognition of Arabic hand sign-based letters and translating them into Arabic speech is proposed in this paper. The proposed system will automatically detect hand sign letters and speaks out the result with the Arabic language with a deep learning model. This system gives 90% accuracy to recognize the Arabic hand sign-based letters which assures it as a highly dependable system. The accuracy can be further improved by using more advanced hand gestures recognizing devices such as Leap Motion or Xbox Kinect. After recognizing the Arabic hand sign-based letters, the outcome will be fed to the text into the speech engine which produces the audio of the Arabic language as an output.

1. Introduction

Language is perceived as a system that comprises of formal signs, symbols, sounds, or gestures that are used for daily communication. Communication can be broadly categorized into four forms; verbal, nonverbal, visual, and written communication. Verbal communication means transferring information either by speaking or through sign language. However, nonverbal communication is the opposite of this, as it involves the usage of language in transferring information using body language, facial expressions, and gestures. Written communication, however, involves conveying information through writing, printing, or typing symbols such as numbers and letters, while visual communication entails conveying information through means such as art, photographs, drawings, charts, sketches, and graphs.

The movement of the arms and hands to communicate, especially with people hearing disability, is referred to as sign language. However, this differs according to people and the region they come from. Therefore, there is no standardization concerning the sign language to follow; for instance, the American, British, Chinese, and Saudi have different sign languages. Since the sign language has become a potential communicating language for the people who are deaf and mute, it is possible to develop an automated system for them to communicate with people who are not deaf and mute.

Sign language is made up of four major manual components that comprise of hands' figure configuration, hands' movement, hands' orientation, and hands' location in relation to the body [1]. There are mainly two procedures that an automated sign-recognition system has, vis-a-vis detecting the features and classifying input data. Many approaches have been put forward for the classification and detection of sign languages for the improvement of the performance of the automated sign language system. The American Sign Language (ASL) is regarded as the sign language that is widely used in many countries such as the USA, Canada, some parts of Mexico, with little modification it is also used

in few other countries in Asia, Africa, and Central America. The research activities on sign languages have also been extensively conducted on English, Asian, and Latin sign languages, while little attention is paid on the Arabic language. This may be because of the nonavailability of a generally accepted database for the Arabic sign language to researchers. So, researchers had to resort to develop datasets themselves which is a tedious task. Specially, there is no Arabic sign language reorganization system that uses comparatively new techniques such as Cognitive Computing, Convolutional Neural Network (CNN), IoT, and Cyberphysical system that are extensively used in many automated systems [2–7]. The cognitive process enables systems to think the same way a human brain thinks without any human operational assistance. The human brain inspires the cognitive ability [8–10]. On the other hand, deep learning is a subset of machine learning in artificial intelligence (AI) that has networks capable of learning unsupervised from data that is unstructured or unlabeled which is also known as deep neural learning or deep neural network [11–15]. In deep learning, CNN is a class of deep neural networks, most commonly applied in the field of computer vision. The vision-based approaches mainly focus on the captured image of gesture and get the primary feature to identify it. This method has been applied in many tasks including super resolution, image classification and semantic segmentation, multimedia systems, and emotion recognition [16–20]. One of the few well-known researchers who have applied CNN is K. Oyedotun and Khashman [21] who used CNN along with Stacked Denoising Autoencoder (SDAE) for recognizing 24 hand gestures of the American Sign Language (ASL) gotten through a public database. On the other hand, the proposal to use Convolutional Neural Network (CNN) for recognizing the Italian sign language was made by Pigou et al. [22]. Whereas Hu et al. had made a proposal for the architecture of hybrid CNN and RNN to capture the temporal properties perfectly for the electromyogram signal which solves the problem of gesture recognition [23]. An incredible CNN model that automatically recognizes the digits based on hand signs and speaks the particular result in Bangla language is explained in [24], which is followed in this work. In [25] as well, there is a proposal of using transfer learning on data collected from several users, while exploiting the use of deep-learning algorithm to learn discriminant characteristics found from large datasets.

There are several other techniques, which are used to recognize the Arabic Sign Language such as a continuous recognition system using the K-nearest neighbor classifier and statistical feature extraction method for the Arabic sign language was proposed by Tubaiz et al. [26]. Unfortunately, the main drawback of the Tubaiz's approach is that the users are required to use an instrumented hand gloves to obtain the particular gesture's information that often causes immense distress to the user. Following this, [27] also proposes an instrumented glove for the development of the Arabic sign language recognition system. The continuous recognition of the Arabic sign language, using the hidden Markov models and spatiotemporal features, was proposed by [28]. Research on translation from the Arabic sign language to text was done by Halawani [29], which can be used on mobile devices. In [30], the automatic recognition using sensor and image approaches are presented for Arabic sign language. [31] also uses two depth sensors to recognize the hand gestures of the Arabic Sign Language (ArSL) words. [32] introduces a dynamic Arabic Sign Language recognition system using Microsoft Kinect which depends on two machine learning algorithms. However, Arabic sign language with this recent CNN approach has been unprecedented in the research domain of sign language. Therefore, this work aims at developing a vision-based system by applying CNN for the recognition of Arabic hand sign-based letters and translating them into Arabic speech. A dataset with 100 images in the training set and 25 images in the test set for each hand sign is also created for 31 letters of Arabic sign language. The suggested system is tested by combining hyperparameters differently to obtain the optimal outcomes with the least training time.

2. Data Preprocessing

Data preprocessing is the first step toward building a working deep learning model. It is used to transform the raw data in a useful and efficient format. Figure 1 shows the flow diagram of data preprocessing.

2.1. Raw Images. Hand sign images are called raw images that are captured using a camera for implementing the proposed system. The images are taken in the following environment:

(i) From different angles

(ii) By changing lighting conditions

(iii) With good quality and in focus

(iv) By changing object size and distance

The objective of creating raw images is to create the dataset for training and testing. Figure 2 shows 31 images for 31 letters of the Arabic Alphabet from the dataset of the proposed system.

2.2. Classifying Images. The proposed system classifies the images into 31 categories for 31 letters of the Arabic Alphabet. One subfolder is used for storing images of one category to implement the system. All subfolders which represent classes are kept together in one main folder named "dataset" in the proposed system.

2.3. Formatting Image. Usually, the hand sign images are unequal and having different background. So, it is required to delete the unnecessary element from the images for getting the hand part. The extracted images are resized to 128×128 pixels and converted to RGB. Figure 3 shows the formatted image of 31 letters of the Arabic Alphabet.

2.4. Dividing Dataset for Training and Testing. For each of the 31 alphabets, there are 125 pictures for each letter. The dataset is broken down into two sets, one for learning set and one for the testing set. A ratio of $80:20$ is used for dividing the dataset into learning and testing set. There are 100

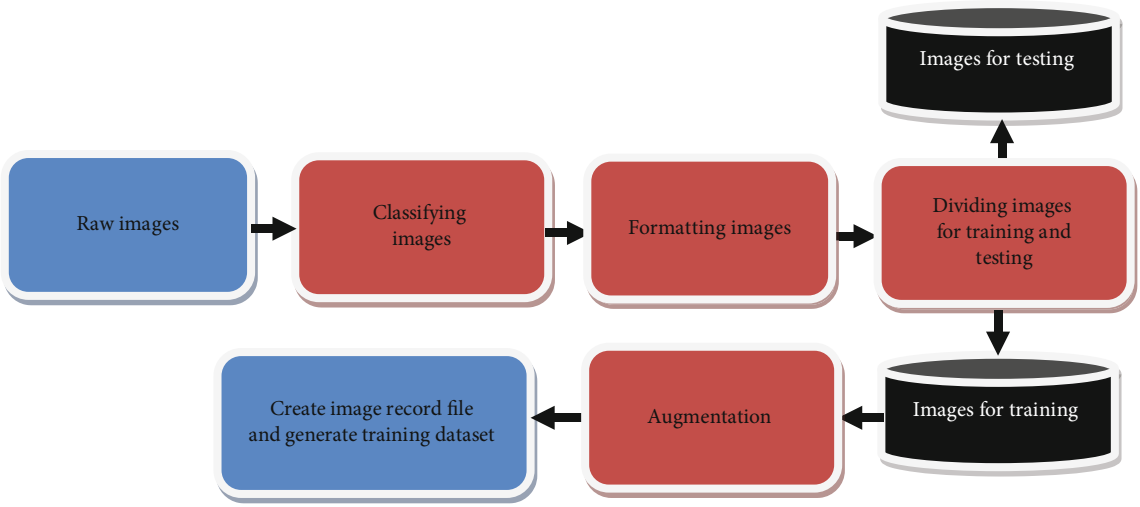


Figure 1: Flow diagram of data preprocessing.

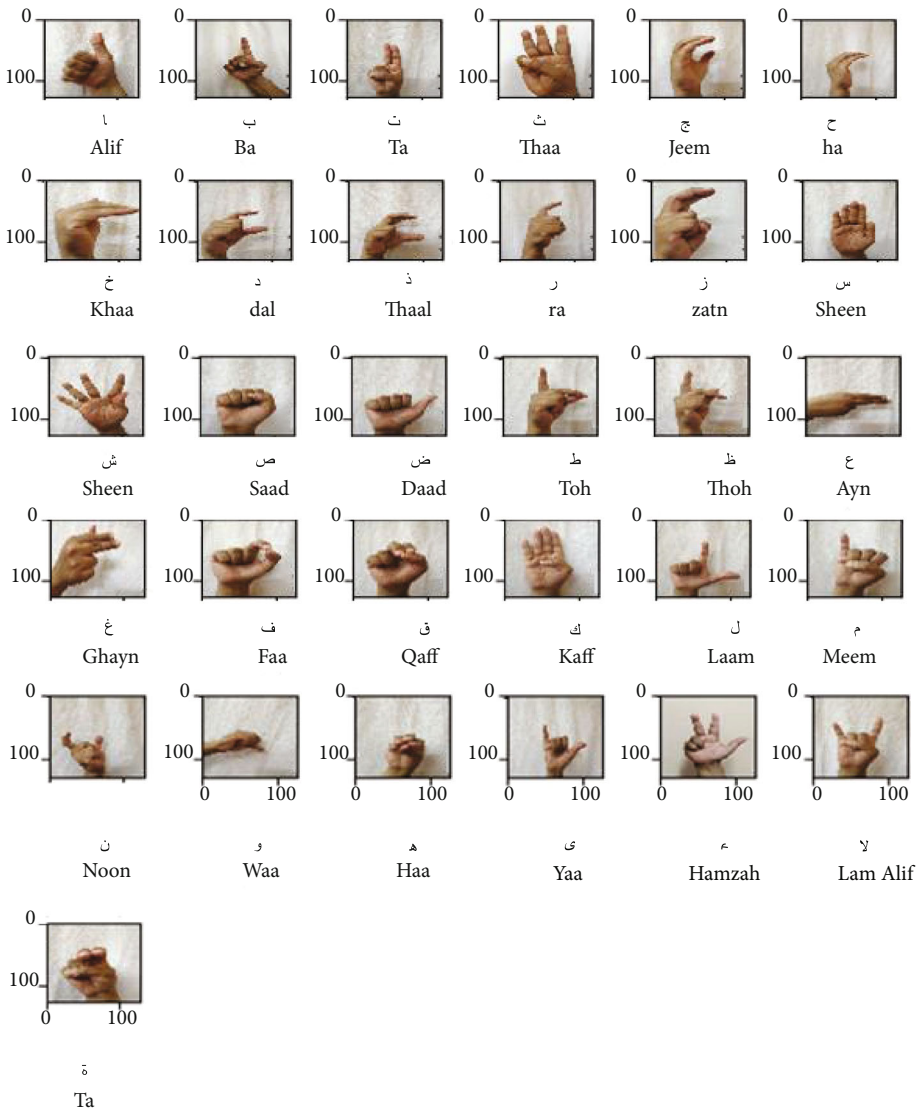


Figure 2: Raw images of 31 letters of the Arabic Alphabet for the proposed system.

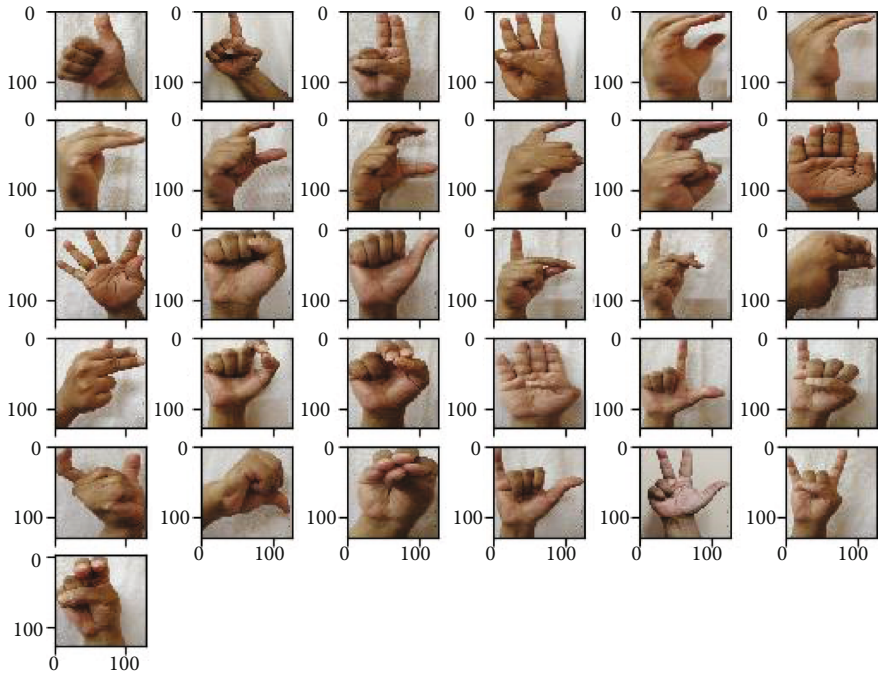


Figure 3: Formatted image of 31 letters of the Arabic Alphabet.

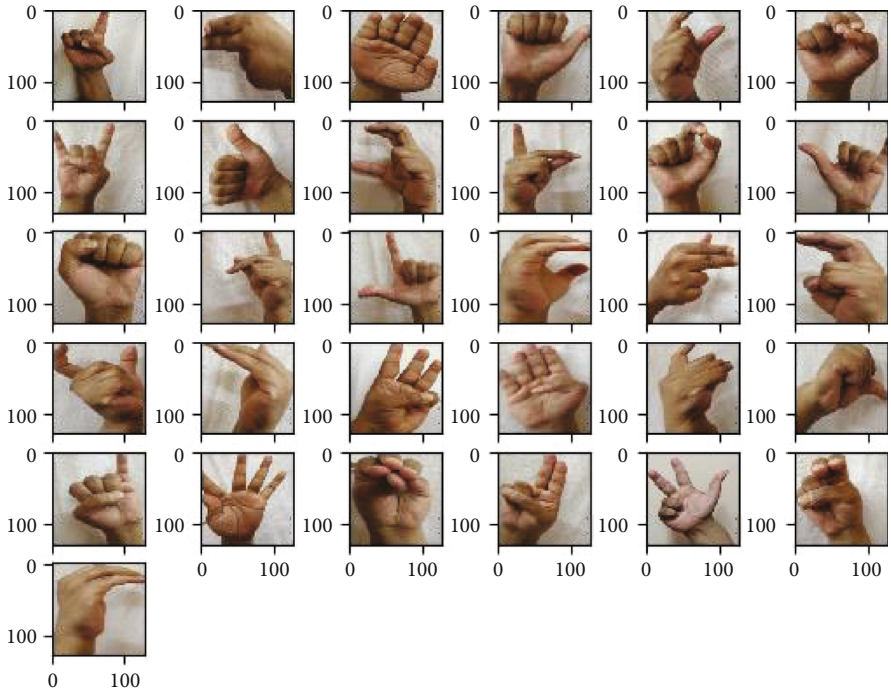


Figure 4: Snapshot of the augmented images of the proposed system.

images in the training set and 25 images in the test set for each hand sign.

2.5. Augmentation. Real-time data is always inconsistent and unpredictable due to a lot of transformations (rotating, moving, and so on). Image augmentation is used to improve deep network performance. It creates images artificially through various processing methods, such as shifts, flips, shear, and rotation. The images of the proposed system are rotated randomly from 0 to 360 degrees using this image augmentation technique. Few images were also sheared randomly with 0.2-degree range and few images were flipped horizontally. Figure 4 shows a snapshot of the augmented images of the proposed system.

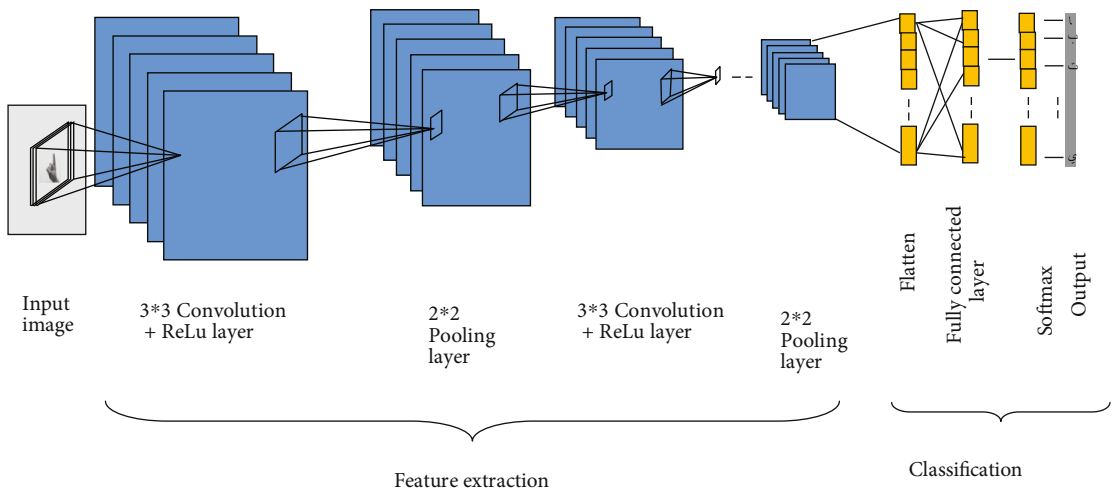


Figure 5: Architecture of Arabic Sign Language Recognition using CNN.

Table 1: Loss and Accuracy with and without Augmentation.

Batch size	Augmentation	Loss	Accuracy
64	False	0.84	83%
	True	0.57	85%
128	False	0.53	86%
	True	0.50	90%

2.6. Create Image Record File and Generate Training Dataset. It is required to create a list of all images which are kept in a different folder to get label and filename information.

3. Architecture

Figure 5 shows the architecture of the Arabic sign language recognition system using CNN. CNN is a system that utilizes perceptron, algorithms in machine learning (ML) in the execution of its functions for analyzing the data. This system falls in the category of artificial neural network (ANN). CNN is mostly applicable in the field of computer vision. It mainly helps in image classification and recognition.

The two components of CNN are feature extraction and classification. Each component has its characteristics that need to be explored. The following sections will explain these components.

3.1. Feature Extraction Part. CNN has various building blocks. However, the major building block of the CNN is the Convolution layer. Convolution layer refers to the mathematical combination of a pair of functions to yield a third function. It is required to do convolution on the input by using a filter or kernel for producing a feature map. The execution of a convolution involves sliding each filter over particular input. At each place, a matrix multiplication is conducted and adds the output onto a particular feature map.

Every image is converted as a 3D matrix by specified width, specified height, and specified depth. The depth is included as a dimension since image (RGB) contains color channels. Numerous convolutions can be performed on input data with different filters, which generate different feature maps. The different feature maps are combined to get the output of the convolution layer. The output is then going through the activation function to generate nonlinear output.

One of the most popular activation function is the Rectified Linear Unit (ReLU) which operates with the computing the function $f(\kappa) = \max(0,\kappa)$. The function shows that the activation is threshold at zero. The ReLU is more reliable and speeds up convergence six times compared to sigmoid and tanh, but it is much fragile during operations. This disadvantage can, however, be overcome by fixing the appropriate learning rate.

Stride refers to the size of a particular step that the convolution filter functions each time. The size of a stride usually considered as 1; it means that the convolution filter moves pixel by pixel. If we increase the size of the particular stride, the filter will slide over the input by a higher interval and therefore has a smaller overlap within the cells.

Because the feature map size is always lesser than the size of the input, we must do something to stop shrinking our feature map. Here, we are intended to use padding.

Now it is required to add zero-value pixels layer to gird particular input by zeros to prevent the feature map from shrinking. Padding also helps in maintaining the spatial dimension constant after doing convolution so that the kernel and stride size matches with the input. So it enhances the performance of the system.

There are three main parameters that need to be adjusted in a convolutional neural network to modify the behavior of a convolutional layer. These parameters are filter size, stride, and padding. It is possible to calculate the output size for any given convolution layer as:

$$\text{Output}_{\text{size}} = \frac{\text{input}_{\text{size}} - \text{filter}_{\text{size}} + 2 * \text{padding}_{\text{size}}}{\text{Stride}_{\text{size}}} + 1, \quad (1)$$

where $\text{output}_{\text{size}}$= the size of the output Convolution layer. $\text{input}_{\text{size}}$= the size of input image. $\text{filter}_{\text{size}}$= the size of filter.

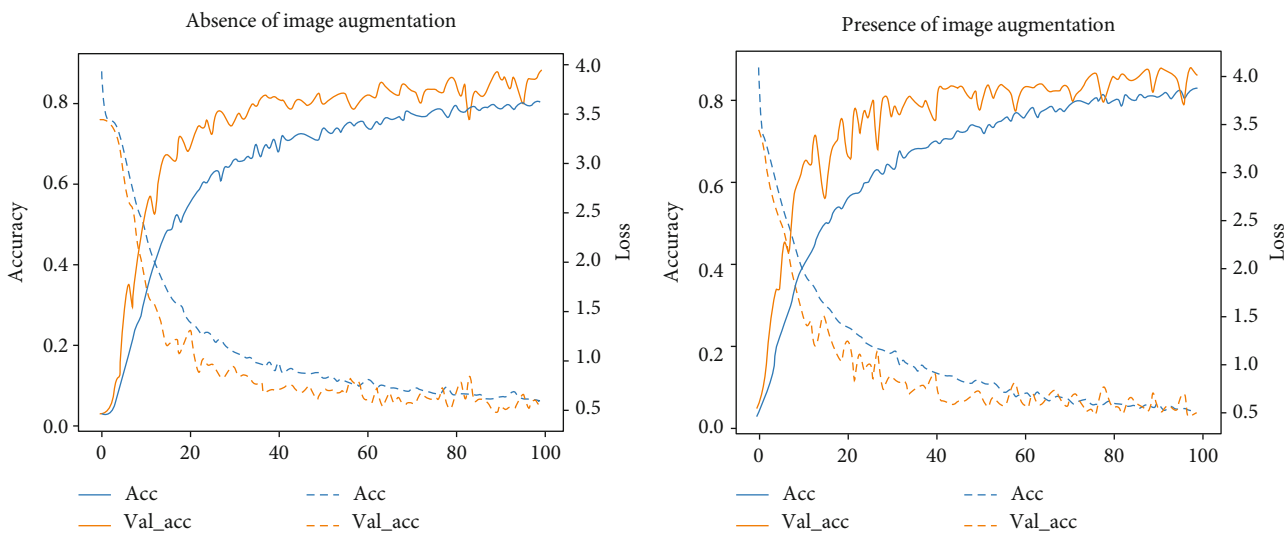


FIGURE 6: Loss and accuracy graph of training and validation in the absence and presence of image augmentation for batch size 128.

$\text{padding}_{\text{amount}}$ = the amount of padding. $\text{stride}_{\text{size}}$= the size of stride.

3.2. Pooling Layer. Naturally, a pooling layer is added in between Convolution layers. However, its main purpose is to constantly decrease the dimensionality and lessen computation with less number of parameters. It also regulates overfitting and reduces the training time. There are several forms of pooling; the most common type is called the max pooling. It uses the highest value in all windows and hence reduces the size of the feature map but keeps the vital information. It is required to specify the window sizes in advance to determine the size of the output volume of the pooling layer; the following formula can be applied.

$$\text{output}_{\text{size}} = \frac{\text{input}_{\text{size}} - \text{filter}_{\text{size}}}{\text{stride}_{\text{size}}} + 1. \quad (2)$$

In all situations, some translation invariance is provided by the pooling layer which indicates that a particular object would be identifiable without regard to where it becomes visible on the frame.

3.3. Classification. The second important component of CNN is classification. The classification consists of a few layers which are fully connected (FC). Neurons in an FC layer own comprehensive connections to each of the activations of the previous layer. The FC layer assists in mapping the representation between the particular input and output. The layer executes its functions by applying the same principles of a regular Neural Network. However, One Dimensional data can only be accepted by an FC layer. For transforming three Dimensional data to one Dimensional data, the flatten function of Python is used to implement the proposed system.

4. Experimental Result and Discussion

The proposed system is tested with 2 convolution layers. Then 2×2 maximum pooling layers follow each convolution layer. The convolution layers have a different structure in the first layer; there are 32 kernels while the second layer has 64 kernels; however, the size of the kernel in both layers is similar 3×3. Each pair of convolution and pooling layer was checked with two different dropout regularization values which were 25% and 50%, respectively. So, this setting allows eliminating one input in every four inputs (25%) and two inputs (50%) from each pair of convolution and pooling layer. The activation function of the fully connected layer uses ReLu and Softmax to decide whether the neuron fire or not. The experimental setting of the proposed model is given in Figure 5.

The system was trained for hundred epochs by RMSProp optimizer with a cost function based on Categorical Cross Entropy because it converged well before 100 epochs so the weights were stored with the system for using in the next phase.

The system presents optimistic test accuracy with minimal loss rates in the next phase (testing phase). The loss rate was further decreased after using augmented images keeping the accuracy almost the same. Each new image in the testing phase was processed before being used in this model. The size of the vector generated from the proposed system is 10, where 1/10 of these values are 1, and all other values are 0 to denote the predicted class value of the given data. Then, the system is linked with its signature step where a hand sign was converted to Arabic speech. This process was completed into two phases. The first phase is the translation from hand sign to Arabic letter with the help of translation API (Google Translator). The generated Arabic Texts will be converted into Arabic speech. In this stage, Google Text To Speech (GTTS) was used.

The system was constructed by different combinations of hyperparameters in order to achieve the best results. The

Table 2: Confusion Matrices with the presence of image augmentation—Ac: Actual Class and Pr: Predicted Class.

PR ⟶ ↓ AC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	2	1	0	12	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0
5	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	12	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	10	4	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	2	0	0	0	1	0	0	0	0	0	0	0	1	5	0	1	0	1	9	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	18	0	0	0	0	0	1	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	10	0	0	1	0	4	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	19	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	18	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	4	0	15	0
31	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	16

results indicated 83 percent accuracy and only 0.84 validation loss for convolution layers of 32 and 64 kernels with 0.25 and 0.5 dropout rate. The system is also tested for convolution layers with batch size 64 and 128. Furthermore, in the presence of Image Augmentation (IA), the accuracy was increased 86 to 90 percent for batch size 128 while the validation loss was decreased 0.53 to 0.50. Table 1 represents these results. It was also found that further addition of the convolution layer was not suitable and hence avoided. Figure 6 presents the graph of loss and accuracy of training and validation in the absence and presence of image augmentation for batch size 128. It is indicated that prior to augmentation, the validation accuracy curve was below the training accuracy and the accuracy for training and loss of validation both are decreased after the implementation of augmentation. The graph is showing that our model is not overfitted or underfitted.

The confusion matrix (CM) presents the performance of the system in terms of correct and wrong classification developed. Therefore, CM of the test predictions in absence and presence of IA is shown in Table 2 and Table 3, respectively.

5. Conclusion

The main objective of this work was to propose a model for the people who have speech disorders to enhance their communication using Arabic sign language and to minimize the implications of signs languages. This model can also be used in hand gesture recognition for human-computer interaction effectively. However, the model is in initial stages but it is still efficient in the correct identification of the hand digits and transferred them into Arabic speech with higher 90% accuracy. In order to further increase the accuracy and quality of the model, more advanced hand gestures recognizing

Table 3: Confusion Matrices in absence of image augmentation—Ac: Actual Class and Pr: Predicted Class.

PR ⟶ ↓ AC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	17	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	13	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	19	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	17	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	1	0	1	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	6	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	1	0	0	10	0	0	0	0	0	2	1	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	3	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	17	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	17	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	14	0	0	0	0	2	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	19	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0
31	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	3	0	0	0	0	0	2	0	0	0	12

devices can be considered such as Leap Motion or Xbox Kinect and also considering to increase the size of the dataset and publish in future work. The proposed system also produces the audio of the Arabic language as an output after recognizing the Arabic hand sign based letters. In spite of this, the proposed tool is found to be successful in addressing the very essential and undervalued social issues and presents an efficient solution for people with hearing disability.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Jouf University, Sakaka, Saudi Arabia, under Grant 40/140.

References

[1] E. Costello, *American Sign Language Dictionary*, Random House, New York, NY, USA, 2008.

[2] Y. Zhang, X. Ma, S. Wan, H. Abbas, and M. Guizani, "Cross-Rec: cross-domain recommendations based on social big data and cognitive computing," *Mobile Networks & Applications*, vol. 23, no. 6, pp. 1610–1623, 2018.

[3] Y. Hao, J. Yang, M. Chen, M. S. Hossain, and M. F. Alhamid, "Emotion-aware video QoE assessment via transfer learning," *IEEE Multimedia*, vol. 26, no. 1, pp. 31–40, 2019.

[4] Y. Qian, M. Chen, J. Chen, M. S. Hossain, and A. Alamri, "Secure enforcement in cognitive internet of vehicles," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1242–1250, 2018.

[5] A. Yassine, S. Singh, M. S. Hossain, and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing," *Future Generation Computer Systems*, vol. 91, pp. 563–573, 2019.

[6] X. Ma, R. Wang, Y. Zhang, C. Jiang, and H. Abbas, "A name disambiguation module for intelligent robotic consultant in industrial internet of things," *Mechanical Systems and Signal Processing*, vol. 136, article 106413, 2020.

[7] M. S. Hossain, M. A. Rahman, and G. Muhammad, "Cyber–physical cloud-oriented multi-sensory smart home framework for elderly people: an energy efficiency perspective," *Journal of Parallel and Distributed Computing*, vol. 103, no. 2017, pp. 11–21, 2017.

[8] K. Lin, C. Li, D. Tian, A. Ghoneim, M. S. Hossain, and S. U. Amin, "Artificial-intelligence-based data analytics for cognitive communication in heterogeneous wireless networks," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 83–89, 2019.

[9] M. S. Hossain and G. Muhammad, "An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 62–68, 2019.

[10] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive internet of things: improving sensitivity and interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58–64, 2019.

[11] X. Chen, L. Zhang, T. Liu, and M. M. Kamruzzaman, "Research on deep learning in the field of mechanical equipment fault diagnosis image quality," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 402–409, 2019.

[12] G. B. Chen, X. Sui, and M. M. Kamruzzaman, "Agricultural remote sensing image cultivated land extraction technology based on deep learning," *Revista de la Facultad de Agronomia de la Universidad del Zulia*, vol. 36, no. 6, pp. 2199–2209, 2019.

[13] P. Yin and M. M. Kamruzzaman, "Animal image retrieval algorithms based on deep neural network," *Revista Cientifica-Facultad de Ciencias Veterinarias*, vol. 29, pp. 188–199, 2019.

[14] G. Chen, Q. Pei, and M. M. Kamruzzaman, "Remote sensing image quality evaluation based on deep support value learning networks," *Signal Processing: Image Communication*, vol. 83, article 115783, 2020.

[15] G. Chen, L. Wang, and M. M. Kamruzzaman, "Spectral classification of ecological spatial polarization SAR image based on target decomposition algorithm and machine learning," *Neural Computing and Applications*, vol. 32, no. 10, pp. 5449–5460, 2020.

[16] B. Kayalibay, G. Jensen, and P. van der Smagt, "CNN-based segmentation of medical imaging data," 2017, http://arxiv.org/abs/1701.03056.

[17] M. S. Hossain and G. Muhammad, "Emotion recognition using secure edge and cloud computing," *Information Sciences*, vol. 504, no. 2019, pp. 589–601, 2019.

[18] M. M. Kamruzzaman, "E-crime management system for future smart city," in *Data Processing Techniques and Applications for Cyber-Physical Systems (DPTA 2019)*, C. Huang, Y. W. Chan, and N. Yen, Eds., vol. 1088 of Advances in Intelligent Systems and Computing, Springer, Singapore, 2020.

[19] Y. Zhang, Y. Qian, D. Wu, M. S. Hossain, A. Ghoneim, and M. Chen, "Emotion-aware multimedia systems security," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 617–624, 2019.

[20] M. S. Hossain, G. Muhammad, W. Abdul, B. Song, and B. B. Gupta, "Cloud-assisted secure video transmission and sharing framework for smart cities," *Future Generation Computer Systems*, vol. 83, pp. 596–606, 2018.

[21] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3941–3951, 2017.

[22] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convolutional neural networks," in *European Conference on Computer Vision*, pp. 572–578, 2015.

[23] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, and W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PLoS One*, vol. 13, no. 10, article e0206049, 2018.

[24] S. Ahmed, M. Islam, J. Hassan et al., "Hand sign to Bangla speech: a deep learning in vision based system for recognizing hand sign digits and generating Bangla speech," 2019, http://arxiv.org/abs/1901.05613.

[25] U. Cote-Allard, C. L. Fall, A. Drouin et al., "Deep learning for electromyographic hand gesture signal classification using transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 760–771, 2019.

[26] N. Tubaiz, T. Shanableh, and K. Assaleh, "Glove-based continuous Arabic sign language recognition in user-dependent mode," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 526–533, 2015.

[27] S. Ai-Buraiky, *Arabic Sign Language Recognition Using an Instrumented Glove, [M.S. thesis]*, King Fahd University of Petroleum & Minerals, Saudi Arabia, 2004.

[28] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, and H. Bajaj, "Continuous Arabic sign language recognition in user dependent mode," *Journal of Intelligent Learning Systems and Applications*, vol. 2, no. 1, pp. 19–27, 2010.

[29] S. Halawani, "Arabic sign language translation system on mobile devices," *IJCSNS International Journal of Computer Science and Network Security*, vol. 8, no. 1, 2008.

[30] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to Arabic sign language recognition," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 551–557, 2014.

[31] M. Almasre and H. Al-Nuaim, "Comparison of four SVM classifiers used with depth sensors to recognize Arabic sign language words," *Computers*, vol. 6, no. 2, p. 20, 2017.

[32] B. Hisham and A. Hamouda, "Supervised learning classifiers for Arabic gestures recognition using Kinect V2," *SN Applied Sciences*, vol. 1, no. 7, 2019.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8865931, 13 pages
https://doi.org/10.1155/2020/8865931

Research Article

A Coupled Grid-Particle Method for Fluid Animation on GPU

Fengquan Zhang,[1,2] **Qiuming Wei**,[1] **and Zhaohui Wu**[2,3]

[1]*School of Information Science and Technology, North China University of Technology, Beijing, China*
[2]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China*
[3]*China Academy of Transportation Sciences, Beijing, China*

Correspondence should be addressed to Fengquan Zhang; fqzhang@ncut.edu.cn

Received 18 March 2020; Accepted 6 May 2020; Published 23 May 2020

Academic Editor: Huimin Lu

In digital production environments, high-quality visual effects play a key role in our mobile device such as game and film. The simulation of fluid animation with free surface is an important area in computer graphic. However, the tracking of fluid surface is a challenging problem because of its instability. In this paper, a coupled grid-particle method for fluid animation surface tracking and detail preserving is proposed. Firstly, based on the nonequilibrium extrapolation method, we design a novel method for reconstructing distribution functions (DFs) of interface grids of lattice Boltzmann method (LBM) and couple the reconstruction method with LBM and volume of fluid (VOF) to track the free surface, which can obtain the accurate surface. Secondly, in order to avoid the loss of details caused by weaknesses in the traditional LBM-VOF method, we design a coupled grid-particle method that not only makes full use of the advantages of the coupled grid-particle method but also realizes the two-way coupling between grid method and particle method. Furthermore, for achieving the real-time requirements of fluid animation, we use GPU parallel computing to accelerate the simulation and use an improved screen space fluid (SSF) rendering method for realistic rendering. The various experiments show that this work can track the fluid surface with high precision and preserve the details of the fluid surface, and it also achieves good real-time performance in large-scale fluid simulation.

1. Introduction

Small-scale features, such as fluid drop, splash, and ripple, show interesting details in realistic physical simulation based on fluid animation. But how to track and preserve them from the numerical model is a challenging problem in computer graphics. In addition, due to the difficulty and high cost of animation production, especially of large-scale natural phenomena, traditional methods are gradually replaced or complemented by virtual reality technology. As one of the important areas in virtual reality, fluid simulation has made remarkable achievements in animation production.

Wen and Ma [1] presented a vorticity preserving lattice Boltzmann method (VPLBM) to simulate high-resolution motion of smoke in real time. Stomakhin et al. [2] obtained the fluid animation based on the material point method, and it was used in the movie Frozen; and Jiang et al. [3] have developed the affine particle in cell method to simulate the animation of water, which has also been used in many movies. The free surface of fluid tracking plays a key technique in many animation simulations. However, the free surface is often limited by the accuracy of solving the advection equation, and errors occur in the visual effect of the animation. In view of the above problems, there are two mainstream solutions: one is to radically solve the problem, that is, to improve the accuracy of solving the advection equation, just like the least-squares volume of fluid (VOF) interface reconstruction algorithm [4]. The other method to solve the problem from visual effect is the coupled grid-particle method. The principle of coupled grid-particle method is to use the grid method with high accuracy to solve the velocity field to evolve the main fluid and then use the particle level set method (PLSM) to track the free surface. The particles beyond the free surface are evolved by the particle method with more details, and the grids and particles are coupled on the free surface [5]. The first solution to improve the accuracy of solving advection equation solution can only reduce but not avoid the error of solving the advection equation.

The second solution to solve the problem from visual effect treats the particles as separate details from the free surface and then uses the particle method to render, focusing on increasing the surface details of fluid. In addition, the accuracy of surface simulation is closely related to the number of particles in PLSM. But too many particles will cause a serious computational burden. Therefore, reasonable preserving and enhancing of the details of the fluid surface become the key to improving the fidelity of visual effects.

This paper focuses on solving the weakness problems using the grid method to improve the fidelity of visual effect. Based on the LBM-VOF method [6], this paper proposes a coupled grid-particle method which can solve the problem of low accuracy of reconstructing DFs of interface grids and abnormal interface grid in the LBM-VOF method. And the contributions of this article are as follows:

(1) This paper proposes a new method for reconstructing DFs of interface grids and couple the reconstruction method with LBM-VOF. This method enhances the accuracy of the reconstruction, to achieve the purpose of preserving the fluid surface detail

(2) Aiming at the problem of anomalous interface grids, this paper proposes a coupled grid-particle method. This method replaces abnormal interface grids with particles and uses the advantage of the particle method to express details to enhance the details of fluid surface. And particles and grids are coupled through a two-way coupling model

(3) For real-time fluid simulation, this paper designs high-performance simulation algorithms and rendering algorithms on GPU

The rest of this paper is organized as follows. After briefly reviewing the related work in Section 2, Section 3 introduces our coupled grid-particle method for fluid animation and describes our improved rendering method on GPU. Finally, experimental results are analyzed in Section 4, and the conclusion and future work are outlined in Section 5.

2. Related Work

In recent years, the simulation of free surface fluid based on LBM has attracted some research interests. Körner et al. [6] proposed the LBM-VOF coupled algorithm for simulating the free surface of the water. However, the accuracy of solving the advection equation is low and the mass exchange error of the interface grid is large, which leads to the suspension of the abnormal interface grid. Thürey and Rüde [7] compared the LBM-VOF coupled algorithm and the coupled algorithm of LBM and the level set method. Their experiment shows that the LBM-VOF coupled algorithm has higher numerical accuracy. Based on the LBM-VOF algorithm, Thurey [8] proposed a method of artificial intervention to solve the suspension problem. Although this method solves the problem of suspension, it leads to mass nonconservation and the evolution of fluid can be greatly interfered with by a human. Aiming at solving the problem of surface reconstruction in LBM-VOF, Janssen and Krafczyk [9] adopted a higher accuracy piecewise linear interface reconstruction (PLIC) method, which fundamentally improved the solution accuracy of the advection equation. However, this method can only improve the accuracy to a certain extent, and it cannot avoid the abnormal interface suspension problem in [8]. Janssen and Krafczyk [10] took advantage of the high parallelism of LBM and implemented the LBM-VOF algorithm on GPGPU, focusing on alleviating the computational burden of fluid simulation through a parallel computing acceleration algorithm. So far, the LBM-VOF method is still the most common free surface tracking algorithm used by studies in LBM, which this paper is focusing on for its improvement. But, the problem of abnormal interface suspension in the LBM-VOF method is urgently needed, but no suitable solution has been proposed yet.

Other than VOF, the LBM method can also be coupled with other surface tracking methods. Zheng et al. [11] built a higher accuracy interface capturing equation and proposed a two-dimensional surface tracking method based on LBM. This method does not require interface reconstruction as needed by the LBM-VOF method, but it costs more computation time than the LBM-VOF method. Kaneda et al. [12] used the coupled level set with VOF (CLSVOF) method to track the two-dimensional free surface of LBM and proposed the LB-CLSVOF model. Although the volume error is reduced, the visual effect still has the problem of interface grid suspension. Wang et al. [5] used PLSM to track the surface of fluid animation based on LBM and SPH to evolve some separated particles. This method coupled LBM and SPH on the free surface. In this method, the particles are regarded as details of the fluid that are separated from the free surface, which are used to increase the free surface details of the fluid animation. In addition, the accuracy of the free surface is closely related to the number of particles in PLSM, and too many particles will cause a serious computational burden.

For detail preserving fluid animation, the research work is relatively sparse. Zhang et al. [13, 14] presented a real-time fluid simulation tool based on the particle method, which can obtain vivid fluid animation. Therefore, in order to obtain realistic fluid animation, efficient detail preserving methods and realistic rendering methods are very important. For particle-based fluid animation, most of the research work focuses on avoiding the loss of details in fluid deformation, such as [15–17]. Wang et al. [18] proposed a new method combined with the Marching Cubes and free surface algorithm to extract the fluid surface and used adaptive surface tension combining the wave-particle equation to show the details of fluid surface. For grid-based fluid animation, Thürey et al. [19] also focuses on fluid control technique instead of detail preserving of fluid surface and proposes a new fluid control technique that used scale-dependent force control to preserve small-scale fluid detail. In addition, some neural network methods such as [20–22] are used in fluid animation recently, which can get fine fluid details.

3. Proposed Method

3.1. The Framework of our Method. As shown in Figure 1, the framework of our method is divided into four parts: LBM,

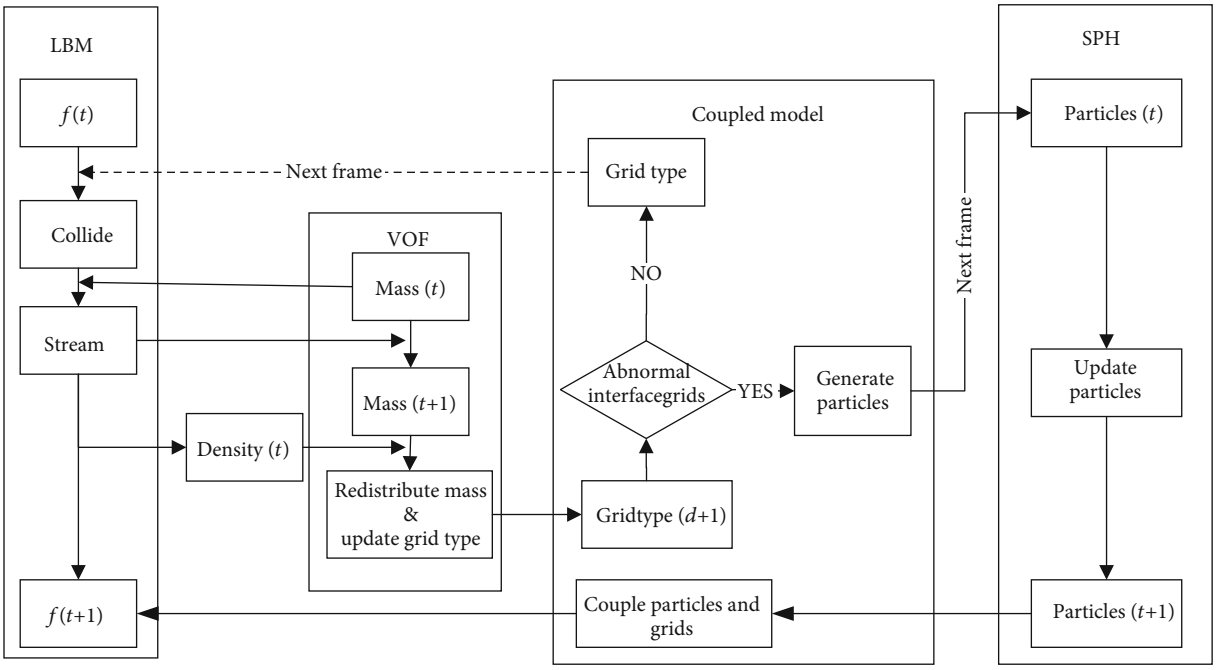


Figure 1: The computational pipeline in our method.

VOF, the coupled model, and SPH. The main part of the fluid adopts the grid-based Lattice Boltzmann method, which can fully utilize the advantages of the grid method in solving the fluid velocity field with high accuracy. In LBM's stream step, a new reconstruction method with second-order accuracy for DFs of interface grid is proposed, which can effectively preserve fluid details. Next, the mass and density between cells are calculated in the stream step, and according to the calculated mass and density, free surface tracking is performed by the VOF method. Then, anomalous interface grids are transformed into particles and particles are coupled with the nonanomalous interface grids through our coupled model. Finally, the particles evolve through the SPH method and use the advantage of the SPH method to express details to enhance surface details.

A new reconstruction method for DFs of interface grid proposed for the LBM-VOF method, but our work is mainly focused on the coupled model. There are three steps in our proposed model. Firstly, the abnormal interface grids in the fluid surface tracked by VOF are extracted while grid types of the nonanomalous interface grids will be used in the next frame. Secondly, the abnormal interface grids are converted to particles that will participate in the evolution in the next frame. Finally, the updated particles of this frame are coupled with the nonanomalous interface grids of the current frame.

Our method is to deal with the abnormal grid generated by the LBM-VOF method through SPH, to achieve detail preserving and enhancing. Moreover, our method can take advantage of the SPH method to depict and enrich the details of fluid surface. In short, our approach combines the advantage of the LBM method which is good at solving the velocity field and the advantage of SPH which is good at depicting surface details.

3.2. LBM-VOF Method for Free Surface Tracking. For free surface tracking, Körner et al. [6] proposed the LBM-VOF coupled algorithm. The fluid solver uses the grid-based LBM, which described the fluid with cellular automation to compute the transport equations. Each cell stores a discrete number of velocities particle. The Bhatnagar-Gross-Krook (BGK) model is applied widely to solve collisions between cells [23], and it can be formulated as

$$f_i(x+e_i, t+\delta_t) - f_i(x,t) = -\frac{1}{\tau}\cdot\left(f_i(x,t) - f_i^{(\mathrm{eq})}(x,t)\right), \quad (1)$$

where e_i is the discrete cell velocity in the direction i, τ is the relaxation time, and $f_i(x,t)$ and $f_i^{(eq)}(x,t)$ are the density DF and the equilibrium density DF in the cell at x at time t, respectively. In 2D simulations, nine directions are usually used in the D2Q9 model, while for 3D simulation, the D3Q19 model with nineteen velocities is the most common one. In our work, we use the velocity vectors as shown in Figure 2.

This paper uses the D3Q19 model and the equilibrium DF, $f_i^{(\mathrm{eq})}(x,t)$, which can be obtained with

$$f_i^{(\mathrm{eq})}(x,t) = \omega_i\rho\left[1 + 3e_i\cdot u + \frac{9}{2}(e_i\cdot u)^2 - \frac{3}{2}(u\cdot u)\right], \quad (2)$$

where $\omega_i = 1/3$ for $i = 0$, $\omega_i = 1/18$ for $i = 1..6$, $\omega_i = 1/36$ for $i = 7..18$, and ρ and u are the density and the velocity of each discrete spatial cell, respectively; they can be obtained by the summation of all DFs for one cell:

$$\rho = \sum_i f_i(x,t), \quad (3)$$

$$\rho u = \sum_i f_i(x,t) \cdot e_i. \tag{4}$$

After the values computed with Equation (1) are stored at DFs for time $t + \delta_t$ (i.e., collide step), each discrete spatial cell needs the DFs of the adjacent cells from the previous time step (i.e., stream step). The collide step and stream step are repeated, and the fluid flows. Additionally, the external forces can be applied in the collide step. This paper uses the GZS model which is of benefit in designing LBM models for fluids exposed to external and internal forces [24]. For boundary conditions, it can be implemented simply by the no-slip boundary condition or the idea extrapolation of the nonequilibrium which has second-order accuracy.

The next step is to update the grid type; the VOF method is used, which is based on tracking the mass of fluid throughout the computational cell. Compared with the level set method, VOF has higher numerical accuracy. Generally, the cells are divided into three types. The cells without fluid are defined as gas phase (G), the partially filled cells are defined as interface (I), and the cells completely are filled as fluid (F). All these cells form a closed layer, as shown in Figure 3.

In other words, "F" cannot be adjacent to "G" directly and only "I" can change to "G" or "F," The interface cells are updated by tracking mass exchange using all neighboring fluid cells and interface cells; the mass exchange is obtained by subtracting the outgoing DFs from the incoming ones during the streaming step:

$$\begin{aligned} \Delta m_i &= \varepsilon \cdot (f_{\text{in}} - f_{\text{out}}) \\ \varepsilon &= \frac{(\varepsilon_1 + \varepsilon_2)}{2} \end{aligned} \tag{5}$$

Here, Δm_i denotes the exchange quantity of mass in direction i. f_{in} and f_{out} are the incoming DF and outgoing DF, while ε_1 and ε_2 are the fraction of the cells that are filled with fluid. As Figure 4 shows, the gas phase does not participate in evolution, so the DFs of "G" are unknown, and its DFs need to be reconstructed.

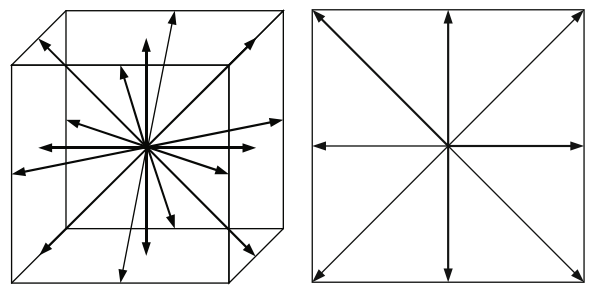
Körner et al. [6] provided a reconstruction method based on the momentum exchange:

$$\bar{f}_{\bar{i}}(x + e_i\delta_t, t + \delta_t) = f_i^{(\text{eq})}(\rho_A, u) + f_{\bar{i}}^{(\text{eq})}(\rho_A, u) - f_i(x,t), \tag{6}$$

where $f_i^{(\text{eq})}(\rho, u)$ is the DF opposite to $f_{\bar{i}}^{(\text{eq})}(\rho_A, u)$, $\bar{f}_{\bar{i}}(x + e_i\delta_t, t + \delta_t)$ is the reconstructed DF of an interface cell which comes from a gas cell, u is the velocity of the interface cell, and ρ_A is an atmospheric pressure parameter. To balance the forces on each side of the interface, the DFs coming from the normal direction of the interface are also reconstructed.

However, this reconstruction method is not accurate. Based on the nonequilibrium extrapolation method, we design a novel method for reconstructing the DFs of the interface grids. The analysis and theory are detailed in Section 3.3.

3.3. A New DF Reconstruction Method for Preserving Surface Detail. The LBM-VOF method for surface tracking has been



(a) D3Q19 model (b) D2Q9 model

Figure 2: The model of velocity vectors.

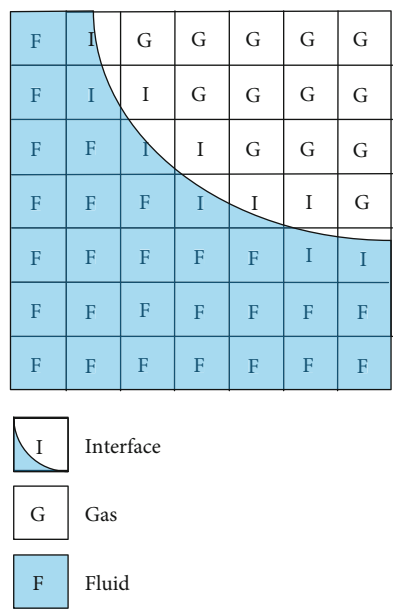


Figure 3: The different grid types required for the free surface.

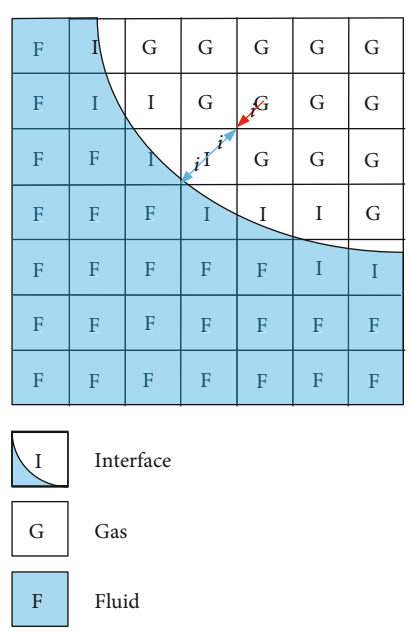


Figure 4: The DFs from the gas phases are unknown.

detailed in Section 3.2, but the poor accuracy of the reconstructed DFs can lead to some artifacts. A typical artifact is that the fluid is initially symmetric, but after a period of evolution, it is clearly asymmetric, which shows that the reconstruction error of DFs is still large.

Guo et al. [25] proposed a nonequilibrium extrapolation method to deal with solid boundaries, whose basic principle is that according to Chapman-Enskog expansion, the DF can be expressed as the sum of the equilibrium function and the nonequilibrium function so that the approximate values can be solved by using the information of adjacent lattice points. This method can be further extended to reconstructing DFs of the interface grids, as shown in Equations (7)–(9):

$$f_{i(\mathrm{Gas})} = f_{i(\mathrm{Gas})}^{(\mathrm{eq})} + \frac{1}{\tau} \cdot f_{i(\mathrm{Gas})}^{(\mathrm{neq})}, \tag{7}$$

$$f_{i(\mathrm{Gas})}^{(\mathrm{eq})} = f_{i}^{(\mathrm{eq})}(\rho_A, u), \tag{8}$$

$$f_{i(\mathrm{Gas})}^{(\mathrm{neq})} = f_{i(\mathrm{Interface})} - f_{i}^{(\mathrm{eq})}(\rho_{\mathrm{Interface}}, u), \tag{9}$$

where the interface grid is in direction i of the gas grid; $f_{i(\mathrm{Gas})}$ is the DF of gas in the direction of i; $f_{i(\mathrm{Gas})}^{(\mathrm{eq})}$ is the equilibrium function of $f_{i(\mathrm{Gas})}$; τ is the relaxation time; $f_{i(\mathrm{Gas})}^{(\mathrm{neq})}$ is the nonequilibrium function of $f_{i(\mathrm{Gas})}$; ρ_A is an atmospheric pressure parameter, $\rho_{\mathrm{Interface}}$ and u are density and velocity of the gas grid, respectively; $f_{i(\mathrm{Interface})}$ is the DF of the interface grid in direction i; and $f_{i}^{(\mathrm{eq})}(\rho_{\mathrm{Interface}}, u)$ is the equilibrium function of the interface grid in direction i which is calculated by $\rho_{\mathrm{Interface}}$ and u.

In the case of low speed, the density and speed of adjacent grids are similar, so the transfer of fluid DF can be considered the transfer of nonequilibrium function. According to Equation (3) and Equation (7), it is obvious that

$$\sum_i f_i^{(\mathrm{neq})} = 0. \tag{10}$$

If only considering direction i and direction $\tilde{i}$, we can obtain

$$f_i^{(\mathrm{neq})} + f_{\tilde{i}}^{(\mathrm{neq})} = 0. \tag{11}$$

According to Equation (7), Equation (8), Equation (9), and Equation (11), we can obtain that

$$f_{i(\mathrm{Gas})} = f_i^{(\mathrm{eq})}(\rho_A, u) + \frac{1}{\tau} \cdot \left(f_{\tilde{i}}^{(\mathrm{eq})}(\rho_{\mathrm{Interface}}, u) - f_{\tilde{i}(\mathrm{Interface})} \right). \tag{12}$$

In this paper, we use Equation (12) instead of Equation (6) to reconstruct DFs of interface grids, as balancing the forces on each side of the interface is no longer needed. We combine this method with LBM-VOF and develop a surface tracking algorithm with higher accuracy than [6, 8].

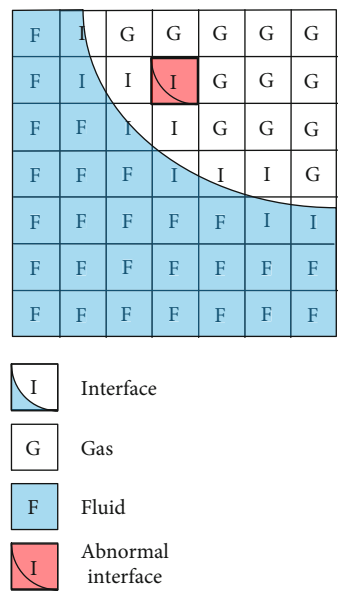


Figure 5: The abnormal interface grid.

3.4. A Coupled Grid-Particle Method for Enhancing Fluid Detail. For free surface tracing, the key is to solve the advection equation, which can be expressed as

$$\frac{\partial \phi}{\partial t} = -u \cdot \nabla \phi. \tag{13}$$

Here, the left side of the formula represents the increment of fluid per unit time, and the right side of the formula represents the amount of fluid inflow per unit time.

The advection equation solution directly determines the accuracy of the free surface of the fluid animation. The key to solving the advection equation is to determine the interface within the grid, which is difficult to solve with high accuracy. As mentioned in Section 3.2, due to the difficulty of accurate determination of the interface within the grid, both the advection equation solution and the mass exchange are inaccurate, resulting in the abnormal suspension of the cells. Typically, when most fluid regions are in equilibrium, there are still many cells suspended in the air. This paper looks at this problem from the perspective of visual effect, that is, solving the cell suspension problem, to provide better visual effects.

Different from [8], this paper does not fully rely on manual intervention but uses a coupled method: the main part of the fluid animation uses LBM with high accuracy to solve the velocity field, and the abnormal grids use SPH to evolve. On the premise of following the physical laws, it can not only solve the cell suspension problem but can also enrich the surface details through particles to a certain extent. This method can be divided into two parts: the generation and evolution of particles and the coupling of LBM and SPH.

3.4.1. The Generation and Evolution of Particles. Based on the LBM-VOF method, the interface grids must be adjacent to the fluid grids. However, as shown in Figure 5, due to the

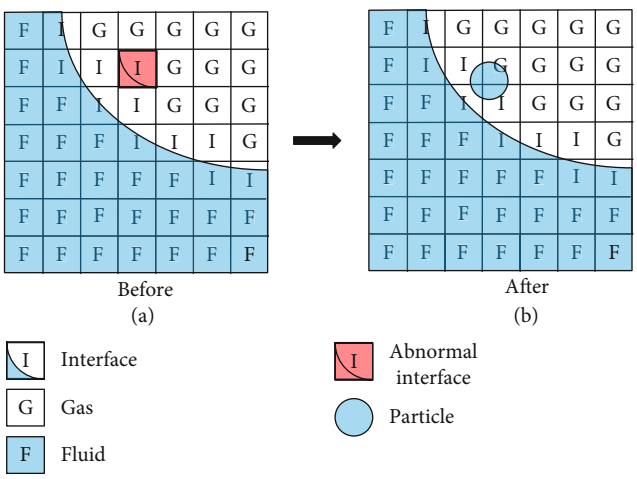


Figure 6: Schematic diagram of particle replacing the abnormal grid: (a) before; (b) after.

low accuracy of mass exchange between interface grids, some interface grids are not adjacent to the fluid grid, and these grids are abnormal interface grids.

To solve the problem of abnormal grids and make abnormal grids flow reasonably, as shown in Figure 6, this paper uses particles instead of these abnormal grids to evolve. There are three steps to generate particles: firstly, generate a particle at the position $\text{pos}_{\text{particle}}$ that is obtained based on the position of the interface grid. Secondly, the physical quantities (mass $\text{mass}_{\text{particle}}$, velocity $v_{particle}$, and density ρ_{particle}) of the particles are calculated according to the physical quantities of the abnormal interface grids, and then, the abnormal interface grids are improved. Finally, in the next frame, the particles evolve in the method of SPH, while the grids evolve with the improved LBM-VOF. The equations used in the three steps are shown as follows:

$$\text{pos}_{\text{particle}} = \text{pos}_{\text{grid}} - (1-\varepsilon)\cdot n, \tag{14}$$

$$\text{mass}_{\text{particle}} = \text{mass}_{\text{grid}}, \tag{15}$$

$$v_{\text{particle}} = v_{\text{grid}}, \tag{16}$$

$$\rho_{\text{particle}} = \varepsilon \cdot m_{\text{grid}}, \tag{17}$$

where pos_{grid} is the position of the abnormal interface grid, ε is the ratio of mass to the density of the abnormal interface grid, and n is the normal vector of the abnormal interface grid. $\text{mass}_{\text{grid}}$, v_{grid}, and ρ_{grid} represent mass, velocity, and density of the abnormal interface grid, respectively.

The SPH method is implemented based on [13, 14]. The density ρ_i, pressure ∇p_i, and viscous force $\mu\nabla\cdot\nabla\mathbf{u}$ of the particles can be calculated by the following three equations:

$$\begin{aligned}
\rho_i &= \sum_{j=1}^{n} m_j W(x_i - x_j, h), \quad i \neq j,\\
\nabla p_i &= -\rho_i \sum_{j=1}^{n}\left(\frac{p_i}{\rho_i^{\,2}} + \frac{p_j}{\rho_j^{\,2}}\right) m_j \nabla W(x_i - x_j, h), \quad i \neq j,\\
\mu\nabla\cdot\nabla\mathbf{u} &= \mu\sum_{j=1}^{n}(\mathbf{u}_j - u_i)\frac{m_j}{\rho_j}\nabla^2 W(x_i - x_j, h), \quad i \neq j,
\end{aligned} \tag{18}$$

where h and $W(x_i - x_j, h)$ are the radius of the smooth kernel and the smooth kernel function of the SPH method, respectively. Specifically, the Poly6 smooth kernel is used in this paper.

3.4.2. The Coupling of LBM and SPH. After dealing with the abnormal grids, the previously evolved particles may coincide with the interface grids, and the physical quantities of the particles need to be transferred to the interface grids in the coincided part. There are six steps in this part:

(1) Calculate the position of the grid pos_{grid} based on where the particle is, defined in the following equation:

$$\text{pos}_{\text{grid}} = \lfloor \text{pos}_{\text{particle}} \rfloor \tag{19}$$

Here, $\text{pos}_{\text{particle}}$ is position of the particle.

(2) As shown in Figure 7, if the grid type of pos_{grid} is interface, the particle needs to be coupled to the grid and execute step (3):

$$\begin{cases} \varepsilon_{\text{interface}} + k\cdot\varepsilon_{\text{particle}} \geq 1 \longrightarrow \text{need coupling},\\ \varepsilon_{\text{interface}} + \varepsilon_{\text{particle}} < 1 \longrightarrow \text{no coupling required},\end{cases}$$
$$k = \begin{cases} 1, \left|\text{pos}_{\text{particle}} - \text{pos}_{\text{grid}}\right| \leq 0.5,\\ 1 - \left(\left|\text{pos}_{\text{particle}} - \text{pos}_{\text{grid}}\right| - 0.5\right), \text{otherwise},\end{cases} \tag{20}$$

where $\varepsilon_{\text{interface}}$ and $\varepsilon_{\text{particle}}$ are volume integrals of the interface grid and particle, respectively

(3) Calculate the new mass of the interface grid m'_{grid} by Equation (21).

$$m'_{\text{grid}} = m_{\text{particle}} + m_{\text{grid}}, \tag{21}$$

where m_{particle} is the mass of the particle and m_{grid} is the mass of the grid

(4) According to the law of conservation of momentum, the new velocity v'_{grid} and density ρ'_{grid} of the

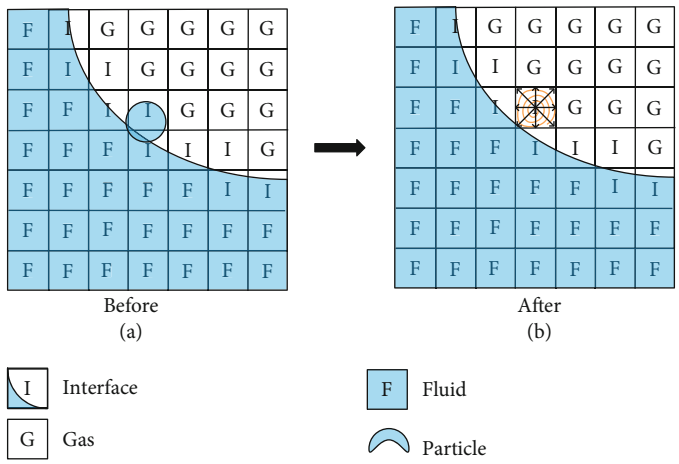


Figure 7: Schematic diagram of coupling of grid and particle: (a) before; (b) after.

Update physical information of particles based on SPH
Update physical information of grids based on improved LBM-VOF
for all grid ∈ interface grids **do**
  if grid ∈ abnormal grids **then**
    Generate particle at position $\boldsymbol{pos_{particle}}$ by using Eq.(14)
    Calculate physical information for the particle by using Eq.(15) - Eq.(17)
    Delete the abnormal grid
  end if
end for
for all particle∈ SPH **do**
  Calculate the grid position $\boldsymbol{pos_{grid}}$ of the particle by using Eq.(19)
  if $\boldsymbol{pos_{grid}}$ ∈ interface grid positions **then**
    Calculate physical information for the grid by using Eq.(21) - Eq.(23)
    Recalculate DF of the grid by using Eq.(1)
    Delete the particle
  end if
end for

Algorithm 1: Coupled grid-particle method for the free surface.

coupled grid can be calculated by Equation (22)–Equation (23):

$$v'_{\text{grid}} = \left(m_{\text{praticle}} v_{\text{praticle}} + m_{\text{grid}} v_{\text{grid}}\right) / m'_{\text{grid}}, \tag{22}$$

$$\frac{m'_{\text{grid}}}{\rho'_{\text{grid}}} = \frac{m_{\text{praticle}}}{\rho_{\text{praticle}}} + \frac{m_{\text{grid}}}{\rho_{\text{grid}}}, \tag{23}$$

where v_{particle} and ρ_{particle} are the velocity of the particle and v_{grid} and ρ_{grid} are the velocity and density of the old grid

(5) Then, according to v'_{grid} and ρ'_{grid} obtained in step (4), the DF of the grid needs to be recalculated by Equation (1)

(6) Delete the particle and repeat step (1)–step (6) until all particles are traversed

The above two parts describe the coupled particle-grid method in this paper. Algorithm 1 summarizes the detailed algorithm flow of this method. This method solves the grid suspension problem from the perspective of visual effect and also enriches the details of the fluid interface to a certain extent. The more detailed experimental analysis will be presented in Section 4.

3.5. Real-Time Fluid Surface Rendering. To meet the real-time requirements of fluid animation, this paper uses GPU to realize the real-time fluid animation simulation. First of all, GPU intuitively supports parallel computing to speed up the whole algorithm. Moreover, the traditional rendering method is abandoned, and the GPU-based screen space fluid (SSF) method is adopted, which does not involve the matching and reconstruction of the fluid surface mesh, and the implementation is based on GPU [26, 27].

In particular, SSF is proposed for particle-based fluid. For rendering grid-based fluid, we improve on the radius and the position of the sprite points. For the radius of the sprite point,

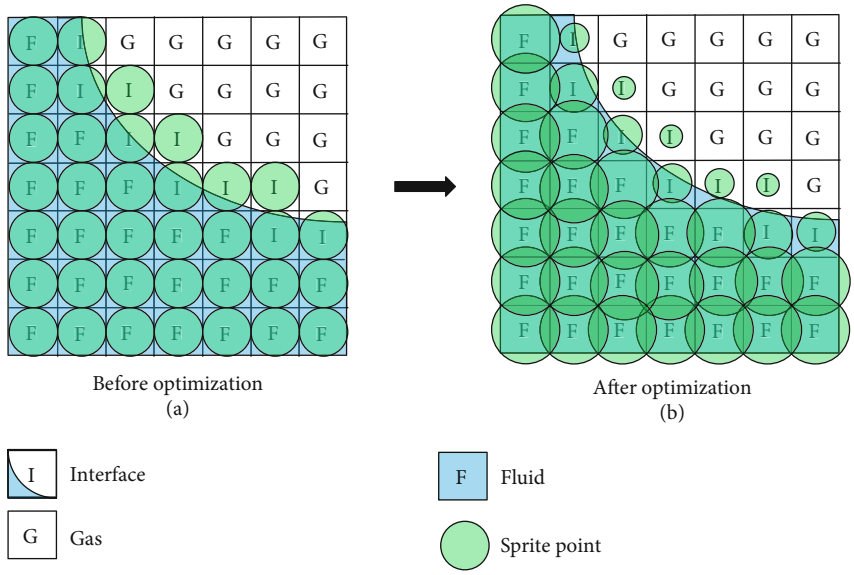


Figure 8: Size optimization of sprite points: (a) before optimization; (b) after optimization.

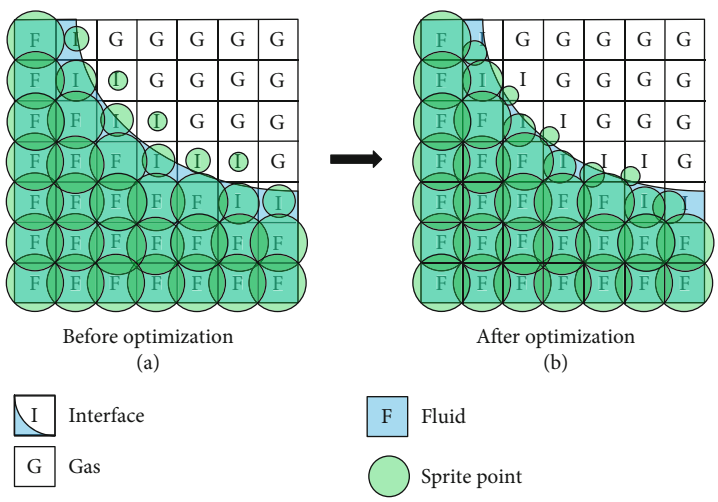


Figure 9: Position optimization of sprite points: (a) before optimization; (b) after optimization.

to make the transformation of the fluid more continuous, the radius of the sprite point is set to K times the density of the grid. K is $\sqrt{2}/2$ in two-dimensional space or $\sqrt{3}/2$ three-dimensional space. As shown in Figure 8, when a grid is filled, the size of the sprite point just covers the grid completely.

For the position of the sprite point, as shown in Figure 9(a), if the sprite points are drawn directly in the center of the grid, there will be obvious gaps between the interface sprite points and the fluid sprite points, which visually represent the discontinuity of fluid flow. For a continuous visual effect, we use Equation (14) to calculate the position of the interface sprite points, and the final schematic diagram is shown in Figure 9(b).

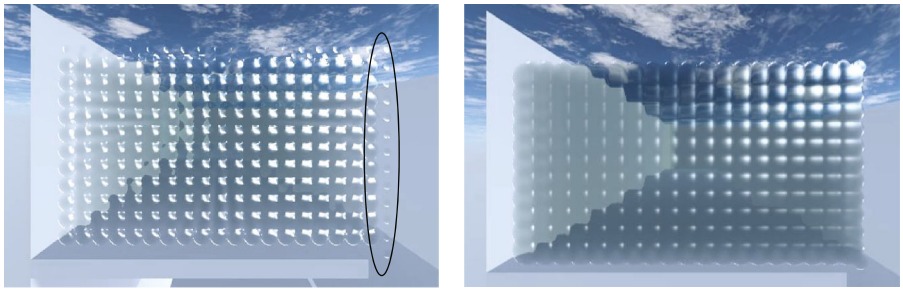
4. Experiment and Analysis

Our experiment is running on a PC with Intel® CPU i5-8300H and NVIDIA GTX 1060 graphics card. All parallel algorithms are implemented by Compute Shader of DirectX, and all fluid rendering algorithms are implemented by CG language.

Our fluid parameter of $1/\tau$ is 1.85, and the fluid parameter of the acceleration of gravity is 0.005, which can be calculated by [8].

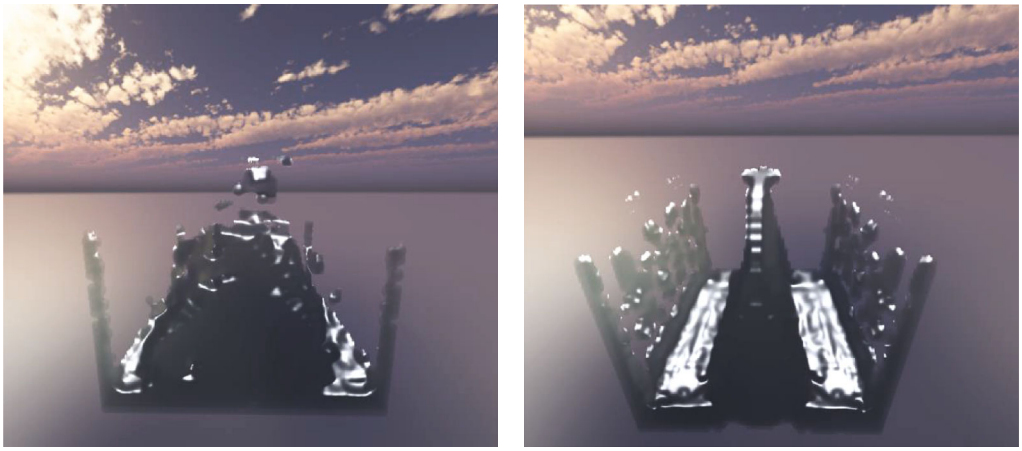
Figure 10 shows the visual effect comparison before and after the optimization of the screen space fluid (SSF) rendering method. Figure 10(a) illustrates the rendering before the optimization, the interface sprite points are separated from the fluid sprite points highlighted in the circle, showing an obvious gap between the interface sprite points and the fluid sprite points. There are even gaps between the fluid sprite points themselves. Figure 10(b) illustrates the rendering after the optimization, the interface sprite points are closer to the fluid sprite points, and the whole surface is smoother.

Figure 11 compares the DF reconstruction method based on [8] and our method at frame 215. Here, $1/\tau$ is 1.85, the



(a) Before optimization (b) After optimization

Figure 10: Comparison of before and after optimization of SSF.



(a) DF reconstruction method based on [8] (b) Our method

Figure 11: High-accuracy reconstruction method of the DFs of the interface grids.

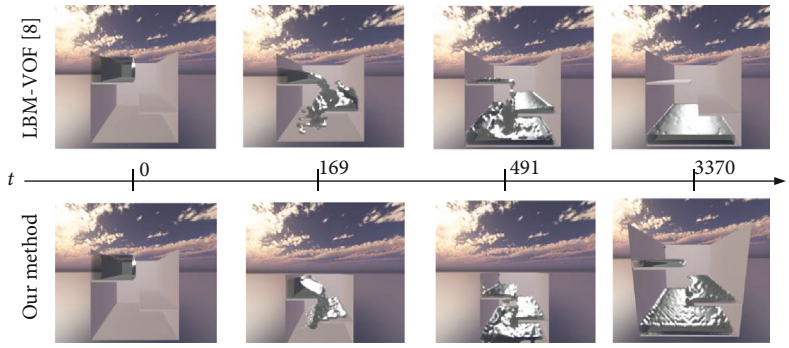


Figure 12: The comparison of the LBM-VOF method [8] and our grid-particle coupled method.

grid size is $40 \times 40 \times 40$, and the fluid parameter of the acceleration of gravity is 0.005 which can be calculated by [8]. Figure 11(a) shows that the fluid no longer remains symmetrical after a period of evolution, while Figure 11(b) shows that most of the fluids on the right and left are symmetrical. The experimental results show that our method has higher accuracy. However, since our method is developed based on the theory of the low-speed model, our method in is only used for low-speed fluid simulation.

A comparative experiment is shown in Figure 12, rendered with a frame rate of 89.1 FPS. The initial condition ($t = 0$) of both methods is the same. Here, the parameters remain the same; we only change the initial positions of the fluid and solid. After the fluids have evolved 169 time steps ($t = 169$), our method shows more interface details as shown in the Figure 12. At the time of $t = 491$, our method shows more interface details that are highlighted by the small circle. The large circle shows the physical phenomenon of the fluid interface affected by the surface tension. This phenomenon is benefited from the high-accuracy DF reconstruction method of interface grids in our method. Finally, when the fluid tends to be more stable ($t = 3370$), the animation based on the LBM-VOF method [8] does not leave any fluid on the obstacles due to the low accuracy of mass exchange and excessive manual intervention, while in our coupled method, there are many fluids left on the obstacles. Specially, our coupled

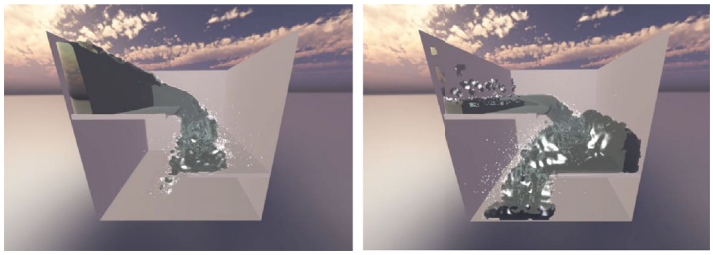


Figure 13: The scene of fluid flowing from the stair.

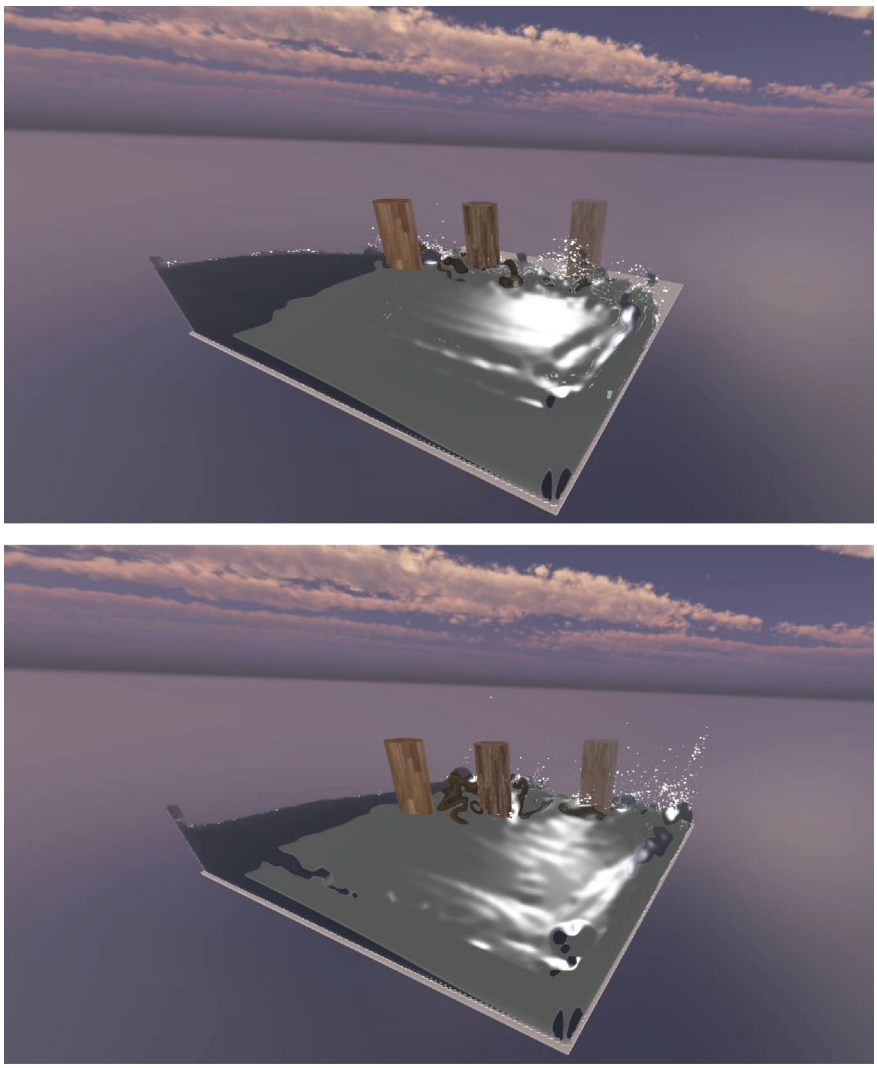


Figure 14: The scene of fluid-solid interaction animation with our method.

model is a simplified physical model, and numerical dissipation inevitably occurs during the numerical interpolation between particles and grids. Therefore, our method mainly solves visual problems and cannot be used for numerical simulation which has high requirements for numerical accuracy.

Figure 13 shows a large-scale fluid animation with an average frame rate of 29.2 FPS, and we increase the grid size to $100 \times 100 \times 100$. Here, it can be seen that our method can not only solve the problem of abnormal interface grid suspension but can also enrich the surface details under the premise of mass conservation. Meanwhile, our method still has good real-time performance in large-scale fluid animation. Similar large-scale fluid simulations are also shown in Figure 14. The number of grid in Figure 13 is 148862, and 96628 in Figure 14.

Table 1 shows the performance of our method in different scenes, and grid size is $100 \times 100 \times 100$. The table shows not only the number of particles at different times but also the average number of frame rate. It can be seen from the

Table 1: The performance of our method in different scenes.

Scene	The number of particles at $t = 100$ s	The number of particles at $t = 200$ s	The number of particles at $t = 400$ s	The number of particles at $t = 800$ s	The averaged time (ms)
Fluid flowing from stair (Figure 13)	3089	12163	23098	13225	30.645
Fluid-solid interaction (Figure 14)	585	5135	1635	786	27.279

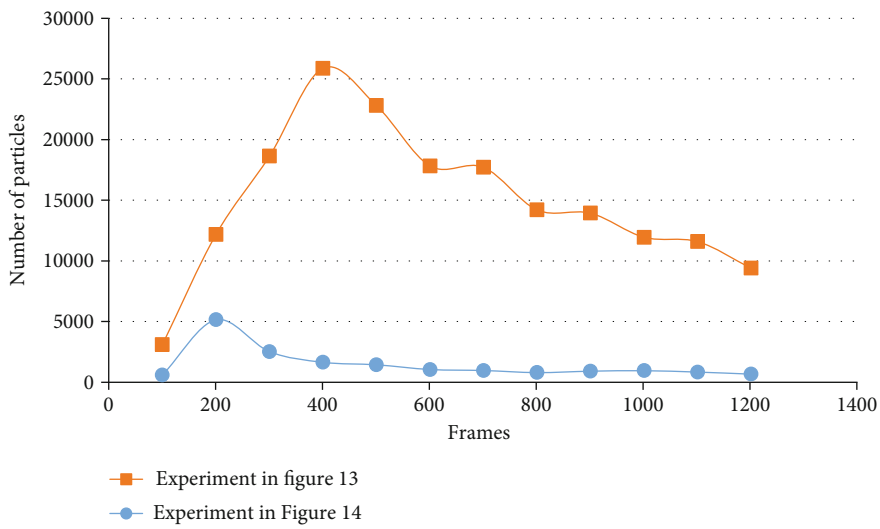


Figure 15: Graph of the change in the number of particles.

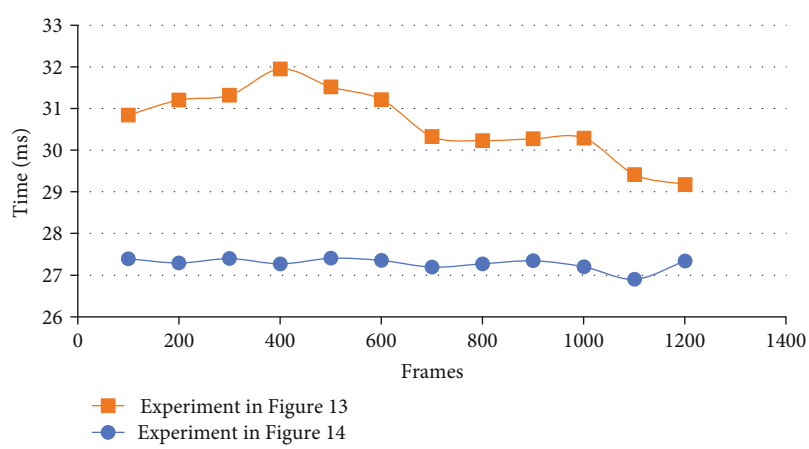


Figure 16: Real-time performance of fluid simulation.

table that the number of particles in Figure 13 is more than the number of particles in Figure 14 at different times, but their average time is about 30 ms. This shows that the number of particles in the optimized algorithm based on GPU does not significantly affect the frame rate of fluid simulation. In fact, the number of particles is orders of magnitude smaller than the number of grids. Finally, more detailed changes in the number of particles during the simulation are shown in Figure 15.

Figure 16 shows the real-time performance data of our GPU algorithm at different times. It is not seen that the performance fluctuation is small during the whole simulation process, and it proves again that the influence of the number of particles on the performance is smaller than that of the fluid grids. And Figure 17 shows more detailed performance data. It can be seen the time in computing takes much more than the rendering. Although the algorithm has reached real-time simulation, there is still a lot of room for optimization in physical computing.

5. Conclusion and Future Work

In this paper, we obtain vivid fluid animation with our proposed method. We improve some key problems in the

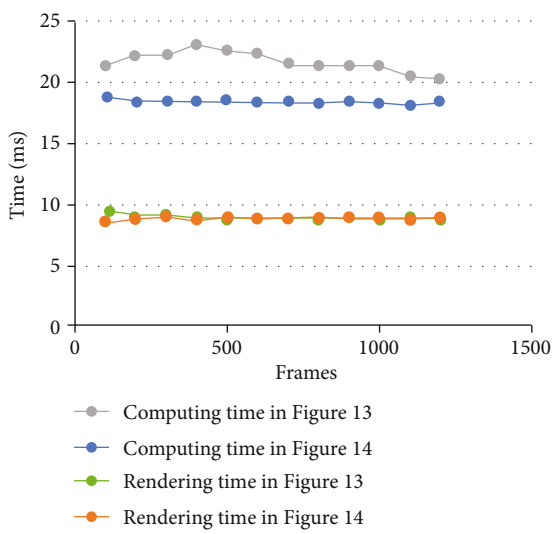


Figure 17: Computing and rendering performance of fluid simulation.

traditional LBM-VOF method. For the first problem of DF reconstruction of the interface grids, we adopted the nonequilibrium extrapolation method which has second-order accuracy to solve this problem and improved this method for reconstructing DFs of interface grids. For the second problem of the abnormal grid suspension caused by the inaccurate mass exchange of interface grids, we first removed the manual intervention which could lead to serious mass nonconservation, and then, we proposed a coupled grid-particle method which combined LBM, VOF, and SPH. Our coupled method not only solved the original problems but also provided enriching details of the fluid interface. Meanwhile, taking advantage of our method's high parallelism, we also used GPU parallel computing to speed up the algorithm and realized real-time rendering.

There are still limitations to overcome. Achieving high-accuracy coupling of grid and particle is still a difficult problem, for example, the transfer of physical properties. Furthermore, when the particles are adjacent to the interface grids, the force between the particles and the interface grids is not considered in this paper. In the future, we will continue to improve the simulation accuracy for more complex fluid animation by extending our method, such as multiple fluid interactions, multiphase flow, and fluid-solid coupling.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is supported by Beijing Natural Science Foundation (No. 4182018, No. 4154067); Beijing Social Science Foundation (No. 18YTC038); Humanities and Social Sciences Fund of the Ministry of Education (No. 19YJC760150); National Natural Science Foundation (No. 61402016); the open funding project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (Nos. VRLAB2018A05 and VRLAB2020B10); Beijing Youth Talent Foundation (No. 2016000026833ZK09); and NCUT Foundation (No. XN018001).

References

[1] J. Wen and H. Ma, "Real-time smoke simulation based on vorticity preserving lattice Boltzmann method," *The Visual Computer*, vol. 35, no. 9, pp. 1279–1292, 2019.

[2] A. Stomakhin, C. Schroeder, L. Chai, J. Teran, and A. Selle, "A material point method for snow simulation," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 1–10, 2013.

[3] C. Jiang, C. Schroeder, A. Selle, J. Teran, and A. Stomakhin, "The affine particle-in-cell method," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 1–10, 2015.

[4] E. Aulisa, S. Manservisi, R. Scardovelli, and S. Zaleski, "Interface reconstruction with least-squares fit and split advection in three-dimensional Cartesian geometry," *Journal of Computational Physics*, vol. 225, no. 2, pp. 2301–2319, 2007.

[5] C.-B. Wang, Q. Zhang, F.-l. Kong, and H. Qin, "Hybrid particle–grid fluid animation with enhanced details," *The Visual Computer*, vol. 29, no. 9, pp. 937–947, 2013.

[6] C. Körner, M. Thies, T. Hofmann, N. Thürey, and U. Rüde, "Lattice Boltzmann model for free surface flow for modeling foaming," *Journal of Statistical Physics*, vol. 121, no. 1-2, pp. 179–196, 2005.

[7] N. Thürey and U. Rüde, "Free surface lattice-Boltzmann fluid simulations with and without level sets," in *Proceedings of the Vision, Modeling, and Visualization Conference 2004 (VMV 2004)*, Stanford, CA, USA, November 2004.

[8] N. Thurey, *Physically based animation of free surface flows with the lattice Boltzmann method, [Ph. D. thesis]*, University of Erlangen, 2007.

[9] C. Janssen and M. Krafczyk, "A lattice Boltzmann approach for free-surface-flow simulations on non-uniform block-structured grids," *Computers & Mathematics with Applications*, vol. 59, no. 7, pp. 2215–2235, 2010.

[10] C. Janssen and M. Krafczyk, "Free surface flow simulations on GPGPUs using the LBM," *Computers & Mathematics with Applications*, vol. 61, no. 12, pp. 3549–3563, 2011.

[11] H. W. Zheng, C. Shu, and Y. T. Chew, "Lattice Boltzmann interface capturing method for incompressible flows," *Physical Review E*, vol. 72, no. 5, article 056705, 2005.

[12] M. Kaneda, T. Ueda, and K. Suga, "Hybrid model of lattice Boltzmann and CLSVOF methods for immiscible two-fluid flow," *Progress in Computational Fluid Dynamics, an International Journal*, vol. 13, no. 3/4, pp. 152–161, 2013.

[13] F. Zhang, Z. Wang, J. Chang, J. Zhang, and F. Tian, "A fast framework construction and visualization method for particle-based fluid," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, Article ID 79, 2017.

[14] F. Zhang, Q. Wei, and L. Xu, "An fast simulation tool for fluid animation in VR application based on GPUs," *Multimedia Tools and Applications*, pp. 1–24, 2019.

[15] F. Gang and L. Shiguang, "Detail-preserving shape deformation in SPH fluid control," *Journal of System Simulation*, vol. 30, no. 6, article 2368, 2018.

[16] G. Feng and S. Liu, "Detail-preserving SPH fluid control with deformation constraints," *Computer Animation and Virtual Worlds*, vol. 29, no. 1, article e1781, 2018.

[17] F. Dagenais, J. Gagnon, and E. Paquette, "Detail-preserving explicit mesh projection and topology matching for particle-based fluids," *Computer Graphics Forum*, vol. 36, no. 8, pp. 444–457, 2017.

[18] C. Wang, Q. Zhang, Z. Zhang, P. Yang, and Z. Xia, "Detail-preserving rendering of free surface fluid with Lattice Boltzmann," in *Transactions on Edutainment VI. Lecture Notes in Computer Science, vol 6758*, Z. Pan, A. D. Cheok, and W. Müller, Eds., pp. 216–226, Springer, Berlin, Heidelberg, 2011.

[19] N. Thürey, R. Keiser, M. Pauly, and U. Rüde, "Detail-preserving fluid control," *Graphical Models*, vol. 71, no. 6, pp. 221–228, 2009.

[20] Y. Xie, E. Franz, M. Chu, and N. Thuerey, "tempoGAN: a temporally coherent, volumetric GAN for super-resolution fluid flow," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–15, 2018.

[21] W. Lu, X. Zhang, H. Lu, and F. Li, "Deep hierarchical encoding model for sentence semantic matching," *Journal of Visual Communication and Image Representation*, article 102794, 2020.

[22] Y. Zhang, W. Lu, W. Ou et al., "Chinese medical question answer selection via hybrid models based on CNN and GRU," *Multimedia Tools and Applications*, 2019.

[23] P. Bhathnagor, E. Gross, and M. Krook, "A model for collision processes in gases," *Physical Review*, vol. 94, no. 3, p. 511, 1954.

[24] Z. Guo, C. Zheng, and B. Shi, "Discrete lattice effects on the forcing term in the lattice Boltzmann method," *Physical Review E*, vol. 65, no. 4, article 046308, 2002.

[25] G. Zhao-Li, Z. Chu-Guang, and S. Bao-Chang, "Non-equilibrium extrapolation method for velocity and pressure boundary conditions in the lattice Boltzmann method," *Chinese Physics*, vol. 11, no. 4, pp. 366–374, 2002.

[26] L. Van, J. Wladimir, S. Green, and M. Sainz, "Screen space fluid rendering with curvature flow," in *I3D '09: Proceedings of the 2009 symposium on Interactive 3D graphics and games*, Boston, MA, USA, February 2009.

[27] N. Truong and C. Yuksel, "A narrow-range filter for screen-space fluid rendering," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 1, no. 1, pp. 1–15, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8822777, 12 pages
https://doi.org/10.1155/2020/8822777

Research Article

Saliency Detection via the Improved Hierarchical Principal Component Analysis Method

Yuantao Chen,[1] Jiajun Tao,[1] Qian Zhang,[2] Kai Yang,[2] Xi Chen,[1] Jie Xiong,[3] Runlong Xia,[4] and Jingbo Xie[4]

[1]*School of Computer and Communication Engineering and Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha 410114, China*
[2]*Department of Electronic Products, Hunan ZOOMLION Intelligent Technology Corporation Limited, Changsha 410005, China*
[3]*Electronics and Information School, Yangtze University, Jingzhou 434023, China*
[4]*Hunan Institute of Scientific and Technical Information, Changsha 410001, China*

Correspondence should be addressed to Yuantao Chen; chenyt@csust.edu.cn

Received 6 April 2020; Revised 14 April 2020; Accepted 17 April 2020; Published 5 May 2020

Academic Editor: Huimin Lu

Aiming at the problems of intensive background noise, low accuracy, and high computational complexity of the current significant object detection methods, the visual saliency detection algorithm based on Hierarchical Principal Component Analysis (HPCA) has been proposed in the paper. Firstly, the original RGB image has been converted to a grayscale image, and the original grayscale image has been divided into eight layers by the bit surface stratification technique. Each image layer contains significant object information matching the layer image features. Secondly, taking the color structure of the original image as the reference image, the grayscale image is reassigned by the grayscale color conversion method, so that the layered image not only reflects the original structural features but also effectively preserves the color feature of the original image. Thirdly, the Principal Component Analysis (PCA) has been performed on the layered image to obtain the structural difference characteristics and color difference characteristics of each layer of the image in the principal component direction. Fourthly, two features are integrated to get the saliency map with high robustness and to further refine our results; the known priors have been incorporated on image organization, which can place the subject of the photograph near the center of the image. Finally, the entropy calculation has been used to determine the optimal image from the layered saliency map; the optimal map has the least background information and most prominently saliency objects than others. The object detection results of the proposed model are closer to the ground truth and take advantages of performance parameters including precision rate (PRE), recall rate (REC), and F-measure (FME). The HPCA model's conclusion can obviously reduce the interference of redundant information and effectively separate the saliency object from the background. At the same time, it had more improved detection accuracy than others.

1. Introduction

The human's visual attention mechanism had enabled humans to do real-time positioning in complex scene images corresponding to the position of important information, in order to determine the priority sequence of different objectives, which can effectively reduce the range of visual processing, thus greatly saving computing resources. Therefore, to study the human's visual attention mechanism and apply it to the research field of computer vision and image processing has great significance. Today, researchers at home and abroad have been widely concerned with the saliency areas of the human's visual attention mechanism based on the detection technology. The method has become an important research topic in the research field of computer vision and has been successfully applied to image cropping, multiple object tracking and recognition, and thumbnail generation.

The researchers in computer vision had often used the bottom-up process to simulate the mechanism of visual attention, which is called the bottom-up saliency model.

For example, Itti et al. [1] had simulated the fusion mechanism of human brain visual cortex neurons to color, brightness, and orientation features, built the visual saliency model based on the principle of center periphery, and effectively detected the saliency area. The calculation process of the model is simple, but the detection of the target area is not accurate. Yang et al. [2] had improved Itti's model [1] based on the graph theory model and proposed the GBVS model. The calculation method is similar to Itti's model [1], and the image's color, brightness, and direction are same. The GBVS model [2] can compute and calculate the saliency map by the Markov random field, and it can detect image saliency from a global perspective. But its drawback is its inefficiency and inability to identify the target contour. Hou and Zhang [3] had put forward to the Spectral Residual (SR) algorithm. Liao et al. [4] had considered that the amplitude spectrum of prior knowledge is subtracted from the amplitude spectrum of the image. The rest is the saliency part of the amplitude spectrum, and then, the target saliency map can be obtained through the transformation in the frequency domain. The algorithm is fast, but the accuracy is difficult to guarantee. The Low-Rank (LR) algorithm had been proposed by Zhou et al. [5]. It is able to extract more notable features from the high-level Apriori to the low-rank framework, but the computation is large, and the saliency map obtained is poor. Generally speaking, the bottom-up saliency methods are mostly basic, faster, and simpler, but the saliency testing results are often represented by dense highlights, so they cannot show the outline of saliency objects.

The visual saliency detection method is the top-down model. According to a specific task, Hou et al. [6] had realized the adjustment of shape, size, feature number, threshold, and so on from the bottom-up testing results. Achanta et al. [7] had proposed the Frequency-Tuned (FT) algorithm. It is the Euclidean distance between the average pixel Gaussian low-pass filtering of each image pixel value in the image. The image's value has saliency value of the image point and formed a kind of measurement method based on the comparison of global saliency detection. The Region Contrast (RC) algorithm had been proposed by Cheng et al. [8], by calculating the saliency value of each partition area and building the saliency map based on local contrast. Dalal and Triggs [9] had proposed a method of human body detection based on the feature of the gradient histogram. The method had used gradient direction histogram information to express human characteristics and extracted human shape information and motion information. It had formed rich feature sets. Through local contrast image features, these top-down models [10–14] were characteristically analyzed; due to the extraction of various features, these models' operation speed is slow and is easily affected by the illumination environment and other objective factors, which makes the object detection accuracy greatly reduced.

In recent years, many researchers had applied such methods as machine learning to saliency detection and made great progress. For example, Yu et al. [15] and Chen et al. [16] had built the Deep Convolution Neural Network (DCNN) model based on the principle of human vision. It has combined with the superpixel clustering method to get the image region features. It can achieve effective saliency detection by learning the features. Zhou and Tang [17] had detected the effectiveness and robustness of machine learning and sparse coding. This method has high robustness, but the operation speed is slow. To this end, Principal Component Analysis (PCA) had been applied to the saliency detection; this method had preserved the efficiency of machine learning [18, 19]. However, when the image background information extracted principal components which had represented saliency goals that cannot be effective, it results in greater detection results with background noise [20, 21].

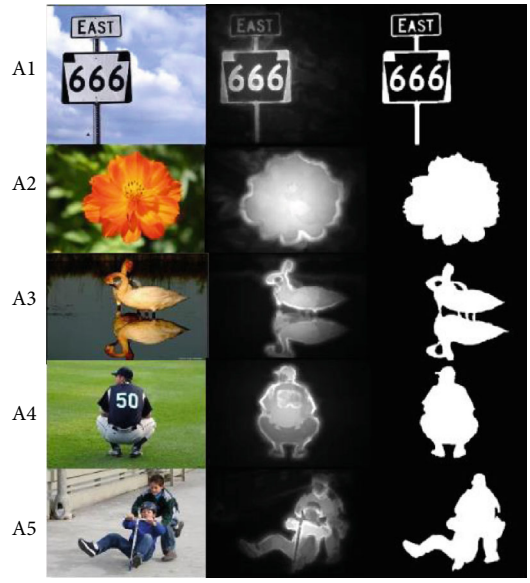


Figure 1: The display results of the proposed algorithm.

In visual saliency detection tasks, due to the complexity of the image, the saliency graph of the single level detection method is not clear [22]. In order to reduce the impact of image complexity, Wang et al. [23] had proposed the Hierarchical Saliency (HS) algorithm. Chen et al. [24] can effectively suppress the interference of background noise to object detection by stratifying the image and calculating the stratified graph.

Based on the above analysis, in order to weaken the impact efficiency of redundant information on the detection results and retention of machine learning, the paper has proposed the saliency object detection algorithm based on the Hierarchical PCA model, using the layered PCA method which divides the image into multilayer images of a lack of background information in different degrees, so that in the process of extracting principal component information in reducing the amount of calculation and weakening the background information of interference to the detection process, it retains the efficiency of machine learning, to increase the robustness of the algorithm. Figure 1 shows the detection results of the proposed algorithm in the paper.

For this paper, the main contributions are as follows: (1) to attenuate the impact of redundant information on testing results and preserve the efficiency of machine learning; (2) to divide the image into multilayer images with different background information which reduces the computational

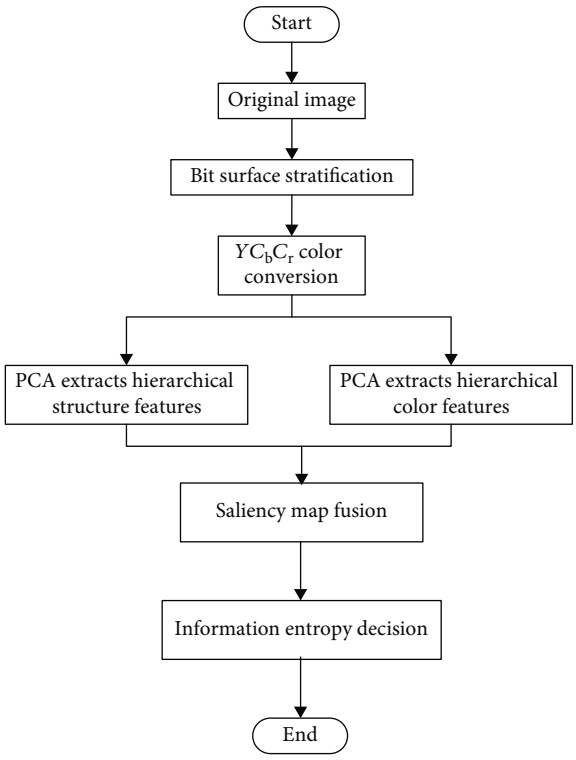


Figure 2: The flowchart of the proposed algorithm.

complexity and reduces the interference of background information on the detection process in the process of extracting the principal component information; and (3) to preserve the efficiency of machine learning and increase the robustness of the proposed algorithm.

Section 2 depicts the proposed algorithm details. Section 3 presents the generating saliency graph with Hierarchical PCA. Section 4 describes the experiments on the MRAS-1000, ASD-1000, and ECSSD-1000 datasets; compares; and analyzes with several methods such as IT, GBVS, SR, LC, HS, BSCA, HDCT, and DCRR. Section 5 summarizes the research work and looks forward to the future research works.

2. The Proposed Algorithm Details

The Hierarchical PCA visual saliency detection algorithm's flowchart is shown in Figure 2. The image information contained in different bit surface layers is quite different, and the eighth image significantly reduces the information contained in the saliency object, so that the significant object area in the image is missing. Other images, to a certain extent, reduce the background information due to the missing bit layers, highlighting the information contained in the saliency objects.

The basic process procedure is as follows: (1) stratification of the original image, using different bit data reconstruction layers which contain an image thus highlights the saliency object information; (2) in order to integrate multiple features, the original color structure is transferred to the gray-level image after stratification, so that each layer of the image has the corresponding color structure corresponding to the original image; (3) PCA is used to extract the structure features and color features; (4) the two distinct features are fused to obtain multiple saliency graphs; and (5) the optimal results are selected through the information entropy. Figure 3 shows an example of the proposed algorithm.

2.1. The Principle of Bit Surface Image Stratification. The eight-bit gray-level image is considered to be composed of eight planes of one bit, each of which contains saliency information that matches it. Four of the high-order bit planes, especially the last two bit planes, contain most of the information of the saliency object. The low-order bit plane contributes to more detailed gray-level details on the image, which means that we can use the saliency information and more bit levels to build the original image, highlighting the proportion of the saliency target in the whole image. Therefore, different bits of information can be used to represent the layered images. The algorithm steps are as follows:

(1) The original image has been converted to a gray image, and it is used as the first layer of the image

(2) The lowest effective bit layer of the first-layer image to zero gets a picture of the image which includes a seven-bit layer as the second-layer image output. The lowest effective bit layer to the second-layer image to zero has the third layer of the image output

(3) The binary data of different bits are converted into decimal pixel values to obtain the multilayer image matching the number of bits

The way of removing binary data and the bit level has been chosen to achieve image stratification. The purpose is to produce images with multiple objects with the dominant target as the main information and to reduce the interference of background information. The operating results are shown in Figure 4.

In Figure 4, they can be seen that different bit planes contain different image information. The eight images obviously reduce the information contained in the saliency objects, so that the saliency object areas in the images are missing. Other images, due to the missing bit layer, also reduce the background information to a certain extent and highlight the information contained in the visual saliency objects.

2.2. The Color Conversion. The image's hierarchy based on the bit surface has been carried out on the basis of the gray-level image. In order to maintain the original color features from layered images, the original image's color structure has been used as the mold to transfer color to the gray-level image after image stratification.

In the color conversion process, the conversional technology has been often used in gray colorful transformation, meaning each image pixel of the black area and white area. They are made of the point of the gray value and sent to the three passages through the implementation of different brightness transformations. It generates the corresponding red value, green value, and blue value, namely, the color image and the pixel corresponding to the color's value, which

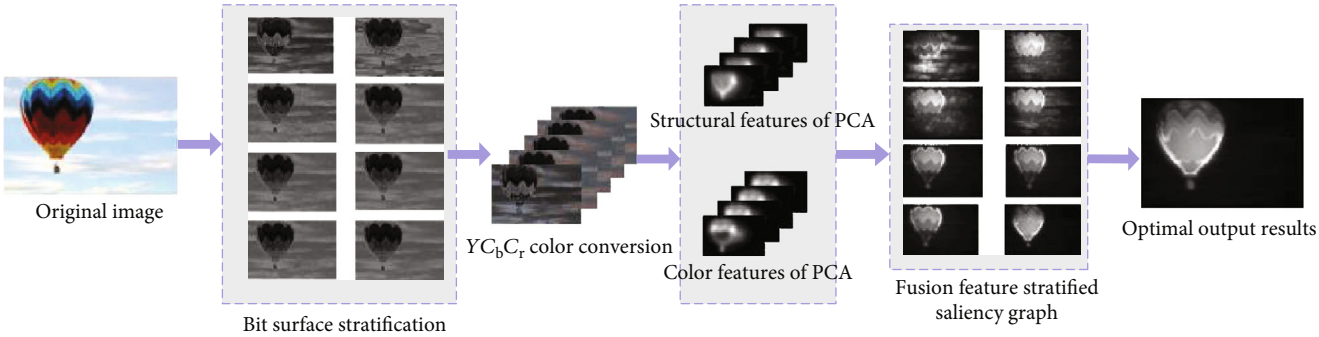


Figure 3: The example of the proposed algorithm.

Input: original image
Output: saliency map
Initialization: adjust the size of the input image
Step 1: bit surface stratification
Step 1.1: the original image has been converted to a gray image, and it is used as the first layer of the image
Step 1.2: the lowest effective bit layer of the first-layer image to zero gets a picture of the image which includes a seven-bit layer as the second-layer image output. The lowest effective bit layer to the second-layer image to zero has the third layer of the image output
Step 1.3: the binary data of different bits are converted into decimal pixel values to obtain the multilayer image matching the number of bits
Step 2: YC_bC_r color conversion
Step 2.1: calculate Mark1 according to formula (3)
Step 2.2: calculate Mark2 according to formula (4)
Step 3: feature extraction using PCA
Step 3.1: calculate $P(p_i)$ according to formula (7)
Step 3.2: calculate the distance $d(p_i, p_a)$ between each image block p_i and average image block p_a according to formula (6)
Step 3.3: calculate the color feature $C(r_i)$ of r_i according to formula (8)
Step 4: saliency map fusion
Step 4.1: calculate the fusion of feature mapping $D(p_i)$ according to formula (9), and limit the range of fusion features to [0, 1] by the normalization method
Step 4.2: combine the fusion feature mapping and Gaussian weight mapping to get the prominent visual saliency map $S(p_i)$ according to formula (10)
Step 5: calculate information entropy $\mathrm{Ens}(x)$ of x according to formula (11), and calculate the saliency graph k_{opt} of the best scale according to formula (12)

Algorithm 1. Proposed algorithm.

can not only retain the mode difference of the object and background of the original image but also enhance the two-color coded target contrast significantly, making the detection more convenient. The implementation step details of color conversion in the paper are as follows.

Firstly, with the original image as the reference image ofI_o(original image) and segmented images as the image to be processed ofI_g (gray image),I_g will be extended to three channels, and the expansion of the image andI_o was transformed into the YC_bC_r space (where Y is the luminance component, C_b is the blue color component, andC_r is the red color component).

Secondly, the maximum and minimum values of every column from the image matrix constituted by I_o are assumed. Assuming that the resolution of the image is $n \times m$, $P(i, j)^n$ $(1 < i < n, 1 < j < m)$ is used to represent the image pixels of the m column; then, the maximum and minimum values of every column in the matrix can be expressed as follows:

$$P\max^m = \max\left(P^m_{1,1}, P^m_{1,2}, \cdots, P^m_{1,n}\right), \tag{1}$$

$$P\min^m = \min\left(P^m_{1,1}, P^m_{1,2}, \cdots, P^m_{1,n}\right). \tag{2}$$

Then, the maximum value of two images, $I_o^{\max}$ and $I_g^{\max}$, and the minimum value of two images, $I_o^{\min}$ and $I_g^{\min}$, are calculated. The two images are normalized to get the reconstructed color image model.

$$\mathrm{Mark1} = \frac{I_g - I_g^{\min}}{I_g^{\max} - I_g^{\min}}, \tag{3}$$

$$\mathrm{Mark2} = \frac{I_o - I_o^{\min}}{I_o^{\max} - I_o^{\min}}. \tag{4}$$

Transfer the colorful map, transfer the image pixel value in Mark2 to the pixel points in the corresponding Mark1, and

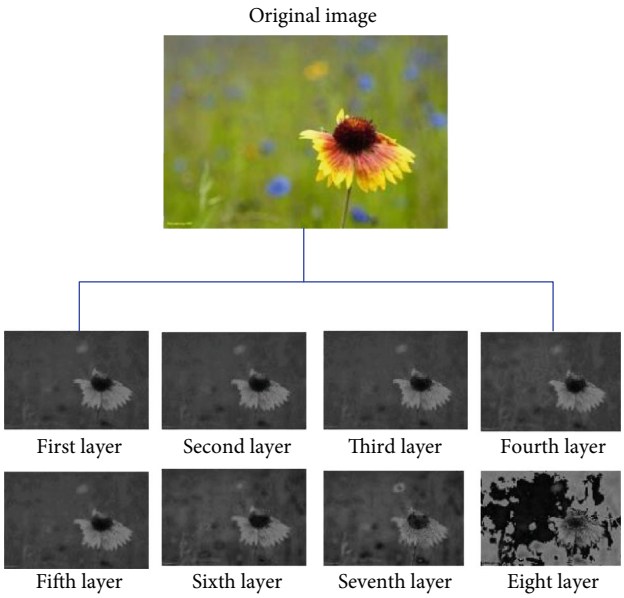


Figure 4: The image stratification results of different layers.

make the hierarchical gray image have the same color structure as the original image. The result of the transformation is shown in Figure 5.

3. The Generating Saliency Graph with Hierarchical PCA

The PCA is the model used to analyze data in the multivariate statistical analysis procedure. It is a way to describe the sample with a small number of features to reduce the dimension of the feature space. The algorithm proposed in this paper is the reconstruction of the saliency map with two features based on the unique structure and color characteristics of pixels near the hierarchical saliency object.

Due to the hierarchical results, the integration model and color pattern of each layer of the image are different from the other images; by section, that image reduces the outstanding target; a layer of background information always exists in the hierarchical image so by calculating each layer image significantly, the results of the output are then found to be most close to the true value. The experimental results are shown in Figure 6, in which Figures 6(a)–6(h) represent the saliency graph of the corresponding stratified images above them, respectively. The specific calculation process of the algorithm is described as follows in Figure 6.

3.1. Extraction of Structural Features. In order to improve the efficiency of structural feature computation, the PCA model based on Wang et al. [23] has been represented in the paper.

Firstly, the layered color image is analyzed by the PCA model, and each layer is divided into 9×9 blocks, and N is the total number of blocks. For a single-layer image, each image block centered on the pixel point i is expressed in p_i, and the average image block p_a can be defined as

$$p_a = \frac{1}{N}\sum_{i=1}^{N} p_i. \tag{5}$$

They can calculate the distance between each image block p_i and average image block p_a, $d(p_i, p_a)$, along principal component direction. Whether an image block has significant structural characteristics is determined based on the distance. Here, the position coordinates of each image block are represented by its central pixel $p_i(i_x, i_y)$, and the position of the average image block is represented by $p_a(a_x, a_y)$. The definition of $d(p_i, p_a)$ is shown as

$$d(p_i, p_a) = \frac{1}{N} \times \sum_{i=1}^{N} p_i \times \sqrt{\frac{(i_x - a_x)^2 + (i_y - a_y)^2}{S^2_{(i,a)}}}. \tag{6}$$

where $S^2_{(i,a)}$ is the variance between the two image blocks.

The rule of judgments is as follows: when $d(p_i, p_a)$ is larger than the threshold of the dataset, the image block is considered to be the saliency area, and the other is a common image block. From the mathematical meaning, the extraction of structural features is attributed to the L_1 norm of p_i in the PCA coordinate system. Therefore, the structural feature $P(p_i)$ is further defined as

$$P(p_i) = \left\|p_i'\right\|_1. \tag{7}$$

In formula (7), p_i' is the coordinate of p_i in the PCA coordinate system. $\|\cdot\|_1$ is the operation symbol of the L_1 norm.

3.2. The Extraction of Color Features. Although the extraction of structural features can find the most unique block in the image, it is not suitable for all images. As shown in Figure 7, the structure characteristics of each sphere are the same, but the colors are different. In this case, they are thinking that the color features are more distinctive. So, the extraction of color features is essential.

Here, two steps are used to detect the color difference of the image block. The first step is to divide each layer of the image into several blocks by using the simple linear iterative clustering superpixel segmentation method and then determine which block has unique color characteristics. In the second step, the sum of distance between the image block and the other image blocks in the YC_bC_r color space is defined as the color difference of the image block. Here, r_i is used to represent the position of i block in the color space, and r_j is used to represent the location of j image block. From the mathematical meaning, the color feature extracted for image block r_i is to calculate its L_2 norm in the PCA coordinate system. Therefore, the color feature$C(r_i)$ ofr_i is defined as

$$P(p_i) = \left\|p_i'\right\|_1. \tag{8}$$

In the upper form, $r_i - r_j$ is represented by the distance between two blocks. $\|\cdot\|_2$ is the operational symbol of L_2

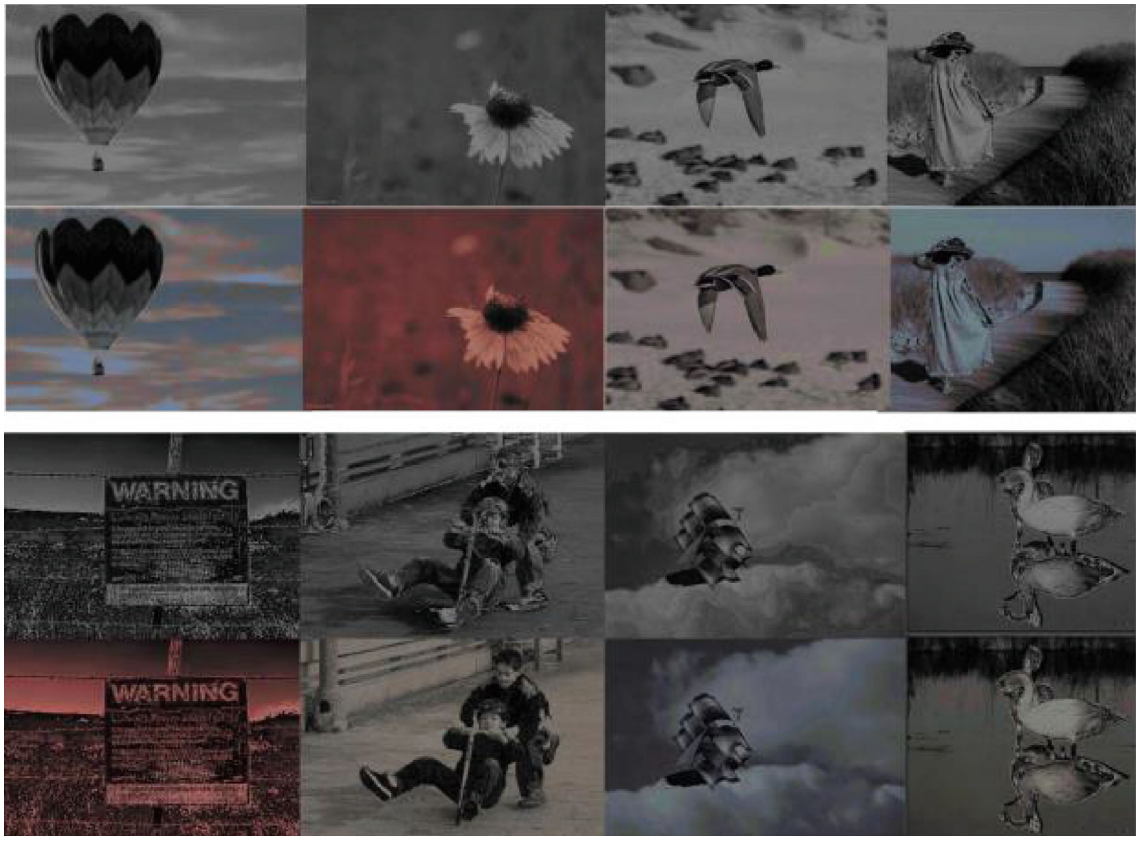


Figure 5: The color conversion effect diagram.

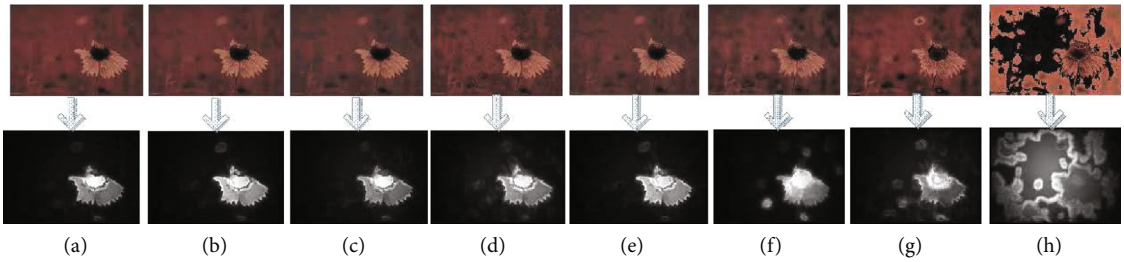


Figure 6: The experimental result diagram.

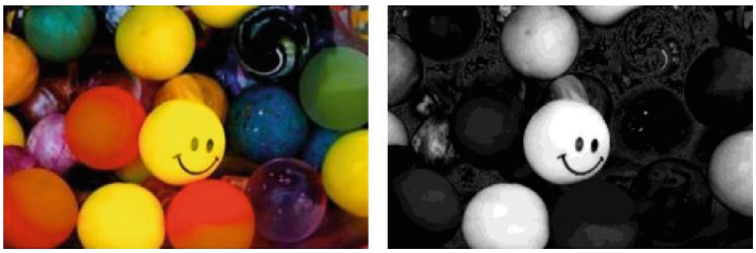


Figure 7: The extraction by the color feature.

norm. M represents the total number of blocks after superpixel segmentation.

3.3. The Saliency Fusion by Structural Features and Color Features. The single image structure feature or color feature cannot effectively characterize all information of the saliency object. In order to obtain accurate and saliency objects, they can combine the structure and color features of each layer of images to detect the saliency regions of different layers of images. Here, they are using the fusion feature to get

$$D(p_i) = P(p_i) \cdot C(p_i). \tag{9}$$

Table 1: Entropy of stratified image information.

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8
B1	6.34	6.39	6.33	6.35	6.36	5.78	6.05	6.10
B2	7.21	6.73	7.03	7.02	7.01	7.03	7.10	7.13
B3	5.78	5.82	5.63	5.56	5.69	5.78	5.79	5.82

Figure 8: The saliency graph using the Hierarchical PCA model.

After that, the fusion features are limited to the $[0, 1]$ range by normalization. Because visual pixels are usually clustered, they usually correspond to objects in real scenes. In order to further modify the saliency models, people usually use the center prior method to put the target area near the center. The center based on the Apriori algorithm usually assumes the target located at the center of the image as a hypothetical condition. By defining the center's prior weight with a peak value-centered Gaussian function, the object saliency in the center of the image is prominently highlighted according to the weight distribution. Here, different target regions are represented by a set of pixels under different thresholds, and the threshold is uniformly distributed in the $[0, 1]$ interval. Therefore, the process of the center prior calculation is as follows:

(1) The image pixel sets of different layers are detected by the fusion of feature mapping $D(p_i)$, and the center of gravity of each threshold result is calculated

(2) The center of gravity places a Gaussian distribution with δ of 10,000, and the corresponding Gaussian weights are calculated for each threshold

(3) The Gaussian distribution with weight of five is added to the image center of each layer to improve the weight of the center position

The Gaussian weight mapping $G(p_i)$ is used to represent the weighted sum of all Gaussian distributions, and different saliency priorities are given according to the difference of weight distribution. Therefore, they can further define the saliency mapping and combine the fusion feature mapping and Gaussian weight mapping to get the prominent visual saliency map $S(p_i)$.

$$S(p_i) = D(p_i) \cdot G(p_i). \tag{10}$$

3.4. The Decision of the Optimal Results. After the above steps, the saliency image corresponding to each layer can be obtained, and the best detection result diagram will be the final output image.

In the information theory, entropy is a relatively basic concept, which is represented by the average amount of information in random events. The information entropy often implies the distribution of the foreground and background noise in the image signal. Generally speaking, if the saliency area of the image is more obvious, it will be more prominent in the whole image performance, and the repeated background area will also inhibit more. Therefore, the saliency region is also gathered in the value of a particular region on the histogram. It provided the small information entropy. The general rule is that the minimum information entropy corresponds to the best saliency graph. For an image signal X, its information entropy is defined by

$$\mathrm{Ens}(x) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p\left(X_{i,j}\right) \log p\left(X_{i,j}\right). \tag{11}$$

In formula (11), $X_{i,j}$ $(i = 1, 2, \cdots, m, j = 1, 2, \cdots, n)$ represents the gray value of image X in line j and line i, and $p(X_{i,j})$ means the probability of occurrence $X_{i,j}$ in image X, and Ens represents the entropy of the image. Then, the saliency graph k_{opt} of the best scale can be expressed as

$$k_{\mathrm{opt}} = \arg\min_{k} \left(\mathrm{Ens}(S_k)\right), \quad k = 1, \cdots, K. \tag{12}$$

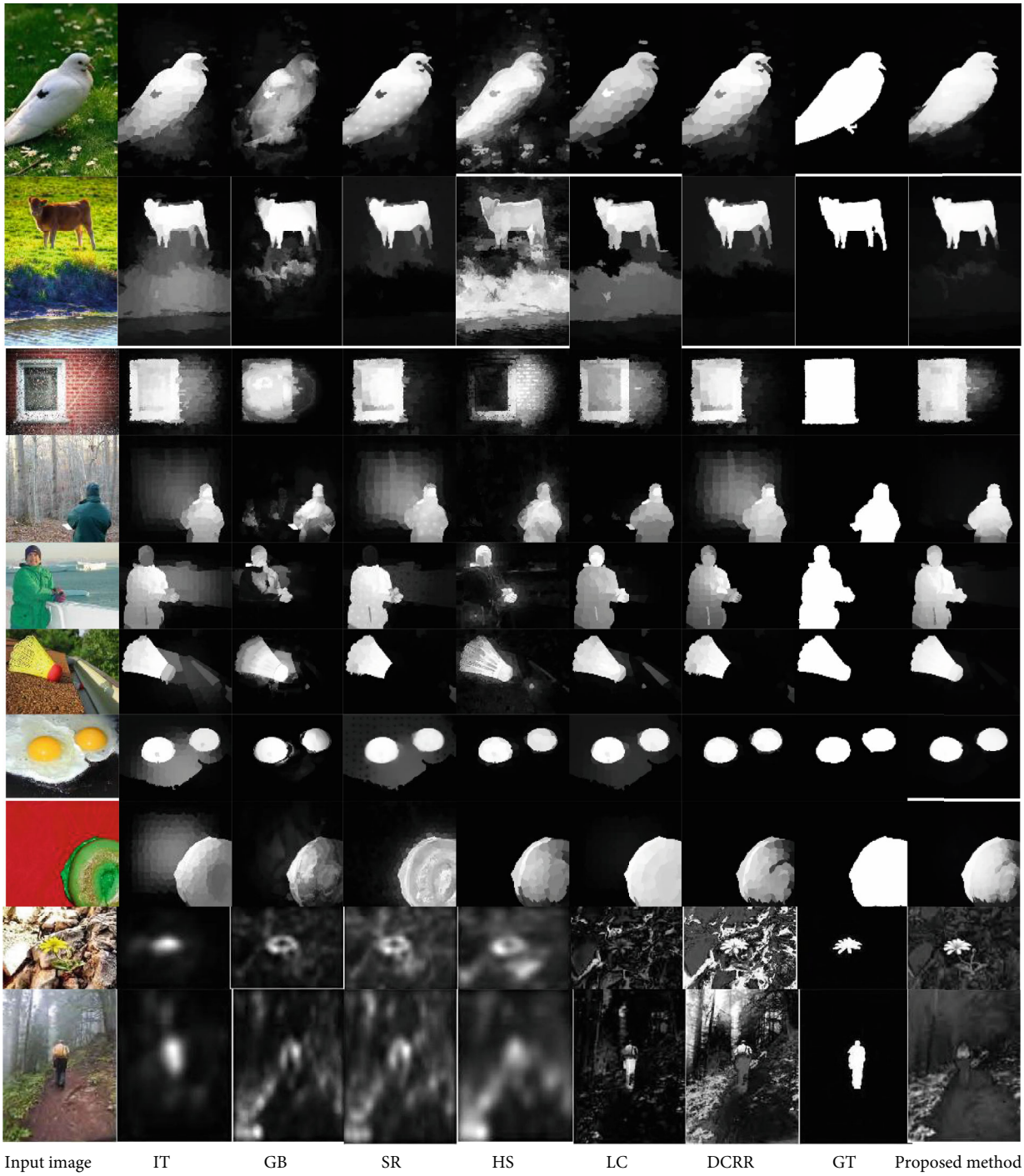


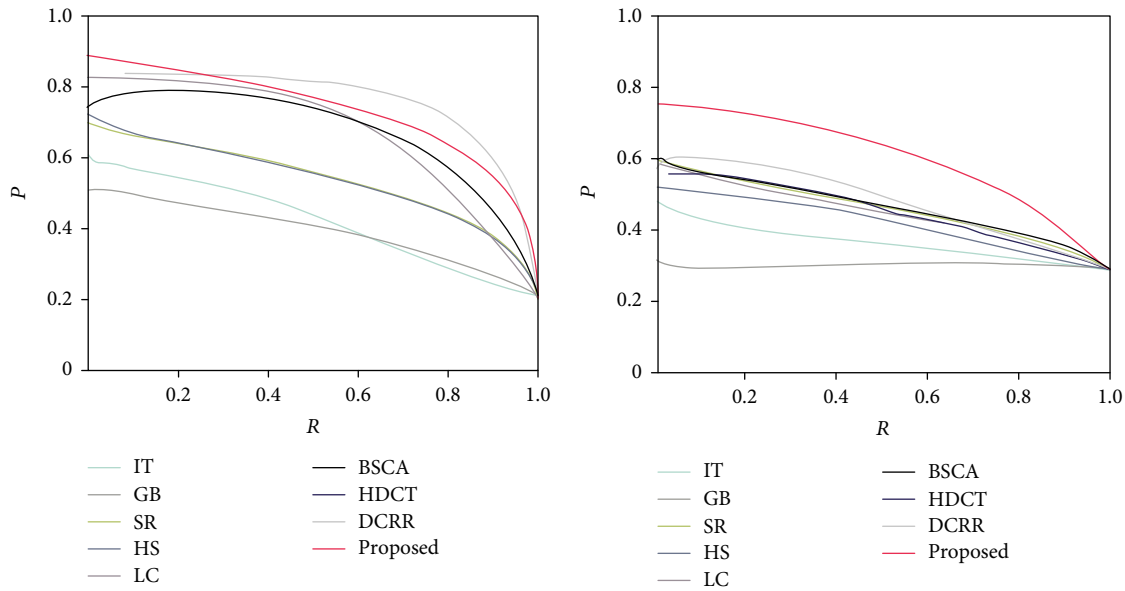
Figure 9: The comparison diagram of algorithm results.

The information entropy is calculated for the multilayer saliency graph after the Hierarchical PCA processing. The information entropy of the stratified image is shown in Table 1.

The data in Table 1 is the information entropy of each image shown in Figure 8. They are using the above information entropy decision rule to decide the eight-level image, select the smallest information entropy image as the output of the optimal result, and get the saliency map with the least background information, which is the final result in Table 1.

4. The Experimental Results and Analysis

The experimental method has used the MATLAB software as the programming platform, and the algorithm is realized on the ThinkPad-E40 laptop. The Hierarchical PCA model in saliency detection is tested on datasets of MRAS-1000, ASD-1000, and ECSSD-1000 and compared with several methods, such as ITTI (IT) [1], GBVS (GB) [2], SR [3], LC [25], HS [23], BSCA [26], HDCT [27], and DCRR [28]. The results of Itti et al. [1] and Yang et al. [2], respectively,



(a) The P-R curve of the ASD dataset

(b) The P-R curve of the ECSSD dataset

Figure 10: The P-R curve of the ASD and ECSSD datasets.

are provided by Hou and Zhang [3], Fang et al. [25], and Wang et al. [23] in each dataset. The CHS [29] had used the original data that is generated on the ECSSD dataset. The result of the visual contrast is shown in Figure 9. In addition, in order to objectively evaluate the detection results, various algorithms are used such as the precision rate (PRE), recall rate (REC), and F-measure (FME) to evaluate the performance. The definitions of PRE, REC, and FME are shown in formulas (13)–(15) [30].

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{13}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{14}$$

$$\mathrm{FME} = 2 \times \frac{\mathrm{PRE} \times \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}}. \tag{15}$$

Among them, TP represents the number of image pixels that detect saliency objects. TN means that the background is correctly divided into the number of pixels in the background class. FP indicates the number of pixels that extract the wrong background. FN means that the saliency object error is divided into the number of pixels in the background class. The AUC indicator is defined as the lower area enclosed by the ROC curve and the coordinate axis, and the maximum value is 1. The larger the AUC, the better the prediction performance of the method on the gaze point of the human eye.

Figure 10 shows the P-R curve [11, 31] of different saliency detection algorithms on three typical common datasets. It can be seen that, because of the high recognition rate of the ECSSD dataset, the accuracy of the HS and LC algorithms is more than 90%, but the precision rate parameter is low. On the ASD dataset, each algorithm reduces the recall to a certain extent, and the precision rate parameter of the algorithms such as IT, GB, and LC is obviously reduced. In the ASD dataset, the recall rate of the GB algorithm is higher, and the HS algorithm has a certain advantage on the accuracy and the F-measure value. In general, the accuracy of the algorithm on different datasets is over 90%, and the average F-measure value is higher than the other algorithms. The detection results are shown as stable and robust.

Figure 11 is a contrast histogram of the F-measure value results of various algorithms shown in Figure 12. It can be seen that due to the high image recognition rate of the ECSSD dataset, the accuracy of the HS and LC algorithms exceeds 90%, but the F-measure value is low. On the more complex MRAS dataset, each algorithm reduces the recall rate to a certain extent, and the F-measure values of IT, GB, LC, and other algorithms are significantly reduced. On the relatively simple ASD dataset, the GB algorithm has a higher recall rate, and the HS algorithm has certain advantages in accuracy and F-measure values. In general, the accuracy of the algorithm in different datasets exceeds 90%, and the average F-measure value is higher than the other algorithms. The detection effect is stable, and the robustness is better than the other solutions.

The results in Table 2 can show the AUC score of each method. It can be seen that our method has the highest AUC score, indicating that our method still has a good detection effect in natural images with complex backgrounds, and can effectively label saliency targets. At the same time, it shows that the method in this paper has higher accuracy and the saliency map obtained is closer to the ground truth.

The calculation speed is an important index for evaluating the superiority of the method. The calculation speed of the method determines whether it can be applied to a real-time system. As a preprocessing process in various image

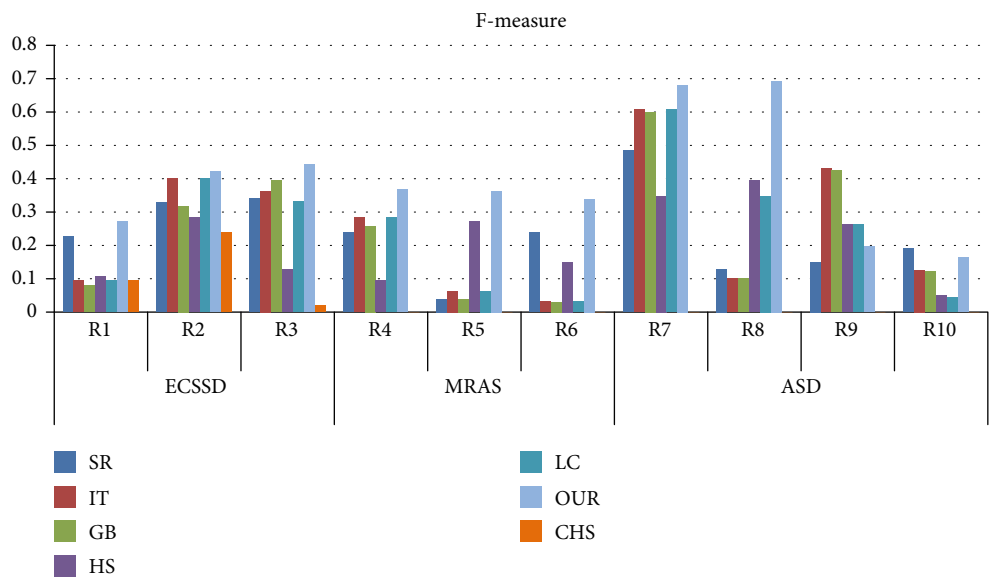
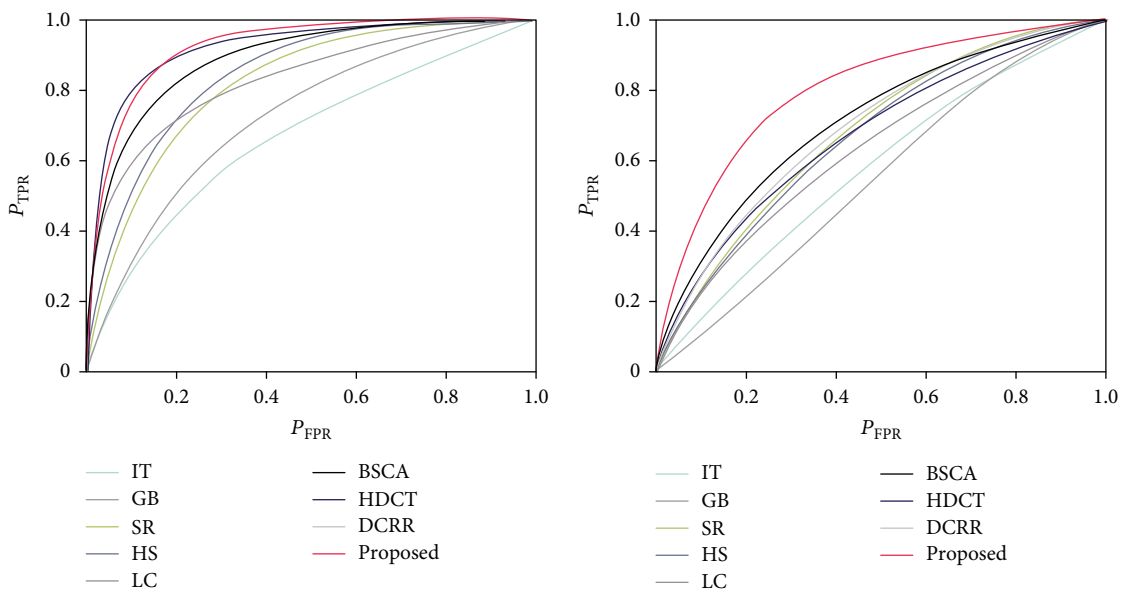


Figure 11: The contrast histogram of the F-measure value.



(a) The ROC curve of the ASD dataset

(b) The ROC curve of the ECSSD dataset

Figure 12: The ROC curve of the ASD and ECSSD datasets.

processing fields, the calculation speed is very important. On the premise that the accuracy of the method meets the expected requirements, the faster the calculation speed, the better the overall performance of the method. The average calculation time of each method is shown in Table 3. The method in this paper has fast calculation speed and can meet the basic application requirements.

5. Conclusions

In this paper, the saliency object detection algorithm based on the Hierarchical PCA model was proposed. The experimental results had shown that the proposed algorithm can reduce the interference of background noise, and the background and target separation has certain advantages in precision, recall, and F-measure parameters, while retaining the excellent characteristics of machine learning methods in order to improve the saliency detection effect. Therefore, the Hierarchical PCA saliency detection is an effective method for object detection under complex backgrounds. The Hierarchical PCA algorithm cannot analyze all the information in the image at the same time. When the objects in the background have the same level of brightness and resolution, it is difficult to extract the complete object information.

Table 2: The AUC value of the ASD and ECSSD datasets.

Method	ASD dataset	ECSSD dataset
IT	0.7252	0.5493
GB	0.8207	0.6681
SR	0.6736	0.5805
HS	0.8232	0.6813
LC	0.8451	0.6954
BSCA	0.8302	0.6755
HDCT	0.8894	0.6954
DCRR	0.9212	0.7091
Proposed	0.9242	0.7990

Table 3: Calculating the time of different methods.

Method	Time (s)
IT	0.2224
GB	0.0163
SR	0.0109
HS	0.0147
LC	0.0288
BSCA	0.0956
HDCT	0.1532
DCRR	0.0752
Proposed	0.0282

Therefore, the future work for the proposed technique is to study the problem of incomplete object information and further improve the information utilization of the whole image to get more accurate and saliency object information.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by grants of the National Natural Science Foundation of China (Nos. 61972056, 61972212, 61402053, and 61981340416), the Open Research Fund of Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation (No. 2015TP1005), the Changsha Science and Technology Planning (Nos. KQ1703018, KQ1706064, KQ1703018-01, and KQ1703018-04), the Research Foundation of Education Bureau of Hunan Province (Nos. 17A007 and 19B005), the Changsha Industrial Science and Technology Commissioner (No. 2017-7), the Natural Science Foundation of Jiangsu Province (No. BK20190089), and the Junior Faculty Development Program Project of Changsha University of Science and Technology (No. 2019QJCZ011).

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173, Portland, OR, USA, June 2013.

[3] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, Minnesota, USA, June 2007.

[4] Z. Liao, R. Zhang, S. He, D. Zeng, J. Wang, and H.-J. Kim, "Deep learning-based data storage for low latency in data center networks," *IEEE Access*, vol. 7, pp. 26411–26417, 2019.

[5] S. Zhou, M. Ke, and P. Luo, "Multi-camera transfer GAN for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 59, no. 2, pp. 393–400, 2019.

[6] X. Hou, J. Harel, and C. Koch, "Image signature: highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[7] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, Miami, Florida, USA, June 2009.

[8] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886–893, San Diego, CA, USA, June 2005.

[10] Y. Chen, J. Xiong, W. Xu, and J. Zuo, "A novel online incremental and decremental learning algorithm based on variable support vector machine," *Cluster Computing*, vol. 22, no. S3, pp. 7435–7445, 2019.

[11] Y. Luo, J. Qin, X. Xiang, Y. Tan, Q. Liu, and L. Xiang, "Coverless real-time image information hiding based on image block matching and dense convolutional network," *Journal of Real-Time Image Processing*, vol. 17, no. 1, pp. 125–135, 2020.

[12] F. Yu, L. Liu, S. Qian et al., "Chaos-based application of a novel multistable 5D memristive hyperchaotic system with coexisting multiple attractors," *Complexity*, vol. 2020, Article ID 8034196, 19 pages, 2020.

[13] F. Yu, Z. Zhang, L. Liu et al., "Secure communication scheme based on a new 5D multistable four-wing memristive hyperchaotic system with disturbance inputs," *Complexity*, vol. 2020, Article ID 5859273, 6 pages, 2020.

[14] G. Sheng, X. Tang, K. Xie, and J. Xiong, "Hydraulic fracturing microseismic first arrival picking method based on non-subsampled shearlet transform and higher-order-statistics," *Journal of Seismic Exploration*, vol. 28, no. 6, pp. 593–618, 2019.

[15] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, "Category-based deep CCA for fine-grained venue discovery from multimodal

data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 1250–1258, 2019.

[16] Y. Chen, J. Wang, S. Liu et al., "Multiscale fast correlation filtering tracking algorithm based on a feature fusion model," *Concurrency and Computation: Practice and Experience*, no. - article e5533, 2019.

[17] L. Zhou and J. Tang, "Fraction-order total variation blind image restoration based on L1-norm," *Applied Mathematical Modelling*, vol. 51, pp. 469–476, 2017.

[18] Y. Yu, S. Tang, F. Raposo, and L. Chen, "Deep cross-modal correlation learning for audio and lyrics in music retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1, pp. 1–16, 2019.

[19] W. Li, H. Xu, H. Li et al., "Complexity and algorithms for superposed data uploading problem in networks with smart devices," *IEEE Internet of Things Journal*, p. 1, 2019.

[20] Z. Liu, Z. Lai, W. Ou, K. Zhang, and R. Zheng, "Structured optimal graph based sparse feature extraction for semi-supervised learning," *Signal Processing*, vol. 170, article 107456, 2020.

[21] H. Lu, Y. Li, S. Mu, D. Wang, H. Kim, and S. Serikawa, "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2315–2322, 2018.

[22] K. Gu, N. Wu, B. Yin, and W. Jia, "Secure data query framework for cloud and fog computing," *IEEE Transactions on Network and Service Management*, vol. 17, no. 1, pp. 332–345, 2020.

[23] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: a discriminative regional feature integration approach," *International Journal of Computer Vision*, vol. 123, no. 2, pp. 251–268, 2017.

[24] Y. Chen, J. Wang, R. Xia, Q. Zhang, Z. Cao, and K. Yang, "The visual object tracking algorithm research based on adaptive combination kernel," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 12, pp. 4855–4867, 2019.

[25] Y. Fang, C. Zhang, J. Li, J. Lei, M. Perreira da Silva, and P. le Callet, "Visual attention modeling for stereoscopic video: a benchmark and computational model," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4684–4696, 2017.

[26] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 110–119, Boston, MA, USA, June 2015.

[27] J. Kim, D. Han, Y. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 883–890, Columbus, OH, USA, June 2014.

[28] H. Lu, D. Wang, Y. Li et al., "CONet: a cognitive ocean network," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 90–96, 2019.

[29] J. Zhang, W. Wang, C. Lu, J. Wang, and A. K. Sangaiah, "Lightweight deep network for traffic sign classification," *Annals of Telecommunications*, 2019.

[30] Y. Chen, J. Wang, X. Chen et al., "Single-image super-resolution algorithm based on structural self-similarity and deformation block features," *IEEE Access*, vol. 7, pp. 58791–58801, 2019.

[31] F. Yu, L. Liu, L. Xiao, K. Li, and S. Cai, "A robust and fixed-time zeroing neural dynamics for computing time-variant nonlinear equation using a novel nonlinear activation function," *Neurocomputing*, vol. 350, pp. 108–116, 2019.