Hindawi
Scientific Programming
Volume 2021, Article ID 6239112, 10 pages
https://doi.org/10.1155/2021/6239112

Research Article

Intelligent Detection Method of English Text in Natural Scenes in Video

Liqin Dai[1] **and ChunHua Chen**[2]

[1]*Jiangxi University of Chinese Medicine, JiangXi 330022, NanChang, China*
[2]*Huai Yin Institute of Technology, JiangSu 223001, HuaiAn, China*

Correspondence should be addressed to Liqin Dai; 20030748@jxutcm.edu.cn

Received 22 October 2021; Revised 5 November 2021; Accepted 9 November 2021; Published 23 November 2021

Academic Editor: Bai Yuan Ding

With the rapid development of Internet technology, breakthroughs have been made in all branches of computer vision. Especially in image detection and target tracking, deep learning techniques such as convolutional neural networks have achieved excellent results. In order to explore the applicability of machine learning technology in the field of video text recognition and extraction, a YOLOv3 network based on multiscale feature transformation and migration fusion is proposed to improve the accuracy of english text detection in natural scenes in video. Firstly, aiming at the problem of multiscale target detection in video key frames, based on the YOLOv3 network, the scale conversion module of STDN algorithm is used to reduce the low-level feature map, and a backbone network with feature reuse is constructed to extract features. Then, the scale conversion module is used to enlarge the high-level feature map, and a feature pyramid network (FPN) is built to predict the target. Finally, the improved YOLOv3 network is verified to extract key text from images. The experimental results show that the improved YOLOv3 network can effectively improve the false detection and missed detection caused by occlusion and small target, and the accuracy of English text extraction is obviously improved.

1. Introduction

As the mainstream part of today's media industry, images and videos are rich in information and easy to understand, which makes them an indispensable part of life. Computer vision analysis is also the key development direction of Internet communication industry at present. For example, character recognition has great application value in many scenes, such as vehicle license plate detection, image-text conversion, image content translation, and image search. However, because the precision of text recognition technology is not ideal, its application scenarios are relatively simple, such as content search in images [1–6].

In image content search scene, the background is simple, the font is single, and some physical features of the image are used to assist, but the error rate is still not ideal. Therefore, most of the current research focuses on text detection and then will consider recognition. Especially in the application of natural scenes, because the background is extremely complex, the font size varies and is affected by special circumstances such as overlapping and pollution of various perspectives, and the difficulty of text detection in natural scenes in videos is much greater than that in ordinary scenes. Image content can be divided into two parts: perceptual content and semantic content [7–9]. The visual part of the image, including the direct visual impression of color, shape, and texture, is the perceptual content. Indirect understanding parts in images, such as objects, words, and events contained in images, are all semantic contents. Among them, words are an important tool for information understanding and communication. Compared with other semantic contents, words are the main content of expressing information and information interaction [10–12]. In addition, the characters are easy to extract and have strong descriptive ability, so how to understand the semantic information of characters in images is an urgent problem to be solved.

This paper focuses on the extraction of English text from natural scenes in video. The meaning of video text

recognition is to extract the text content in the video and then recognize it by the recognition system and finally get the text content. Video is essentially a sequence stream composed of a series of images, and the words in the frame images can express the contents in a short time. One kind of text is the text in the natural scene of the image [13–15], such as the license plate number and bus stop sign text in the image [16], and the other kind of text is artificially added, such as movie subtitles, advertising information, and medical image analysis text [17]. Therefore, all the words should be extracted except the repeated words within a short time delay. The final result of the recognition system can only be determined if the images and characters have good detection performance, so the text detection is the key research content of this paper.

2. Literature Review

In the aspect of text detection, traditional methods are mainly based on the characteristics of text coherence and single color. For example, Minetto et al. [18] proposed an improved image text detection algorithm. Firstly, this algorithm extracts three edge images with different colors by using edge detection operators obtained from three directions and then applies common operations in morphology to these three edge images in turn to obtain different connected domain images. Finally, the three connected graphs are AND-operated, and the noise is filtered, so as to obtain the text region. Yin et al. [19] also adopted a similar method, combined with the morphological method, and used the OSTU algorithm to obtain adaptive threshold, so as to obtain a clean and clear binary image. According to the characteristics of different gray trends of characters and backgrounds, Risnumawan et al. [20] proposed a video character location method based on gradient discrete cosine transform algorithm. In this method, each frame is divided into $n \times n$ blocks, and the discrete cosine transform coefficients of each block are calculated. The amplitude obtained by the gradient operator is used as the block strength for smooth filtering and morphological processing. Finally, the image is projected horizontally and vertically, and the candidate text regions are extracted by wavelet transform and unsupervised clustering. Zhuge and Lu [21] used the model based on level set function to realize text segmentation with small color difference between target and background and large text groove and solved the problem of difficult parameter selection in variational model through optimization calculation.

With the popularization of computers and the great improvement of computer technology, especially the powerful parallel computing capability of GPU and the Big Data resources in the era of mobile Internet, CNN continues to develop. Text detection also follows this trend, turning to the method based on CNN technology, and the effect is greatly improved, which is a step closer to the application of real scenes. Alqhtani et al. [22] trained CNN to detect characters in text, calculated the confidence of pixel blocks, and adopted the nonmaximum suppression method when locating text lines. Then, a similar transcription process is adopted, the classification scores of each character are inquired, and the best words are selected for matching. Tong et al. [23] integrated text detection and recognition into an end-to-end network. RPN (ridge polynomial network) is used for detection, and bilinear sampling is used to unify the text regions into highly consistent variable-length feature sequences, and then recurrent neural network (RNN) is used for recognition. You Only Look Once (YOLO) is a target recognition and location algorithm based on the deep convolution neural network [24]. Its most obvious advantage is its fast detection speed [25], which is especially suitable for the real-time detection system, which is the fundamental reason why YOLOv3 network is selected in this paper. YOLO has been continuously improved on the basis of the original version and has developed to v3 version [26]. However, conventional object detection methods in the visual field (SSD, YOLO, faster-RCNN, etc.) are not ideal when directly applied to English text detection tasks. The main reasons are as follows: compared with conventional objects, the length of text lines and the ratio of length to width vary widely. Therefore, after analyzing YOLO series networks, we propose a new multiscale feature fusion method, which improves the performance of YOLOv3 networks.

In order to construct features with multiscale characteristics and rich expressive ability, we introduce a feature scale transformation and migration fusion method to improve the traditional YOLOv3 network. Different from the existing multiscale feature fusion model by scaling single-channel features, we achieve the purpose of feature scale reduction and enlargement by splitting and combining multichannel feature maps. At the same time, we migrate and fuse the converted features in the backbone network and construct the backbone network with reduced feature scale and FPN prediction network with enlarged feature scale, which achieves better detection results than YOLOv3 network when detecting occluded and smaller targets.

3. English Text Detection Based on Improved YOLOv3 Network

3.1. Convolutional Neural Network. At present, the principles of target detection algorithms are generally divided into two methods: region division and position regression. For example, fast-RCNN method can obtain high accuracy, but its running speed is slow. The latter, for example, SSD and YOLO pursue the real-time performance of the algorithm but can also obtain acceptable detection results. Among these methods, YOLO has become a widely used and efficient algorithm because of its fast speed and high precision. YOLO is a target recognition and location algorithm based on the deep convolution neural network, and its most obvious advantage is its fast detection speed, which is especially suitable for the real-time detection system, which is the fundamental reason why YOLOv3 network is selected in this paper.

Convolution neural network extracts features by convolution operation on local "receptive field" [27], and it is mainly used in image processing-related problems.

CNN is a kind of feedforward neural network with deep structure. Firstly, the image is input at the input layer and then calculated by the convolution layer, pooling layer, and nonlinear activation function, and the semantic information of high-level abstraction is gradually extracted from the image. This is the "feedforward operation" of the convolutional neural network. Finally, for the fully connected layer, all the features extracted from the previous network are connected for prediction, and the difference between the detected value and the true labeled value of the network is calculated. The loss is propagated back to the first convolution layer from the fully connected layer by the gradient descent method, so that all the parameters of the network are updated, and the whole network model converges after several rounds of training. The shallow features extracted by convolution network are shown in Figure 1.

3.2. Feature Fusion. Fusion of convolution features of different scales is an important means to improve the performance of target detection. The low-level features have higher resolution and contain more position and detail information, but because there are few convolution layers, they have less semantic information and more noise. High-level features have stronger semantic information, but their resolution is very small, and their ability to perceive details is poor. How to fuse them efficiently? This section introduces the feature fusion methods of upsampling, deconvolution, and scale transformation in detail.

3.2.1. Upsampling. The upsampling method is to directly interpolate the original feature map [28] to enlarge the feature map, and interpolation is the most common and practical method. On the basis of the original feature map, interpolation algorithm is used to insert new pixel values between the original pixel positions, thus enlarging the scale of the feature map smoothly.

Bilinear Interpolation is an interpolation method for two-dimensional features, which is an extension of linear interpolation algorithm and is widely used in the field of image processing. The purpose of bilinear interpolation method is to calculate the position element value by using the existing values of the target point in two vertical directions in the original feature map, and its main purpose is to jointly determine a linear interpolation in two directions. Schematic diagram of bilinear interpolation is shown in Figure 2.

In Figure 2, the data points with known values are marked in red and represented by the letter Q; mark the data points to be interpolated as green and use the letter P to represent them; mark the data points in the middle transition in blue, which is indicated by the letter R. The values of R_1 and R_2 can be obtained by formulas (1) and (2), respectively.

$$f(R_1) = \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}), \tag{1}$$

$$f(R_2) = \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}). \tag{2}$$

The value of the target point P is inserted by linear interpolation in the y direction.

$$f(P) = \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2). \tag{3}$$

Bilinear interpolation method uses four points in two vertical directions in the original image to calculate the target pixel value for interpolation.

3.2.2. Deconvolution. The formula for calculating a single deconvolution layer is as follows:

$$\sum_{k=1}^{K_1} z_k^i \oplus f_{k,c} = y_c^i. \tag{4}$$

In this layer, an image y^i composed of feature images $y_1^i, \ldots, y_{K_0}^i$ of K_0 color channels is used as input. Each channel c of the image can be expressed as the linear sum of K_1 potential feature maps and convolution kernel.

The deconvolution layer makes the potential feature graph z_k^i sparse by introducing regularization terms. The total loss function of is composed as follows:

$$C_1\left(y^i\right) = \frac{\lambda}{2} \sum_{c=1}^{K_0} \left\| \sum_{k=1}^{K_1} z_k^i \oplus f_{k,c} - y_c^i \right\|_2^2 + \sum_{k=1}^{K_1} \left| z_k^i \right|^p, \tag{5}$$

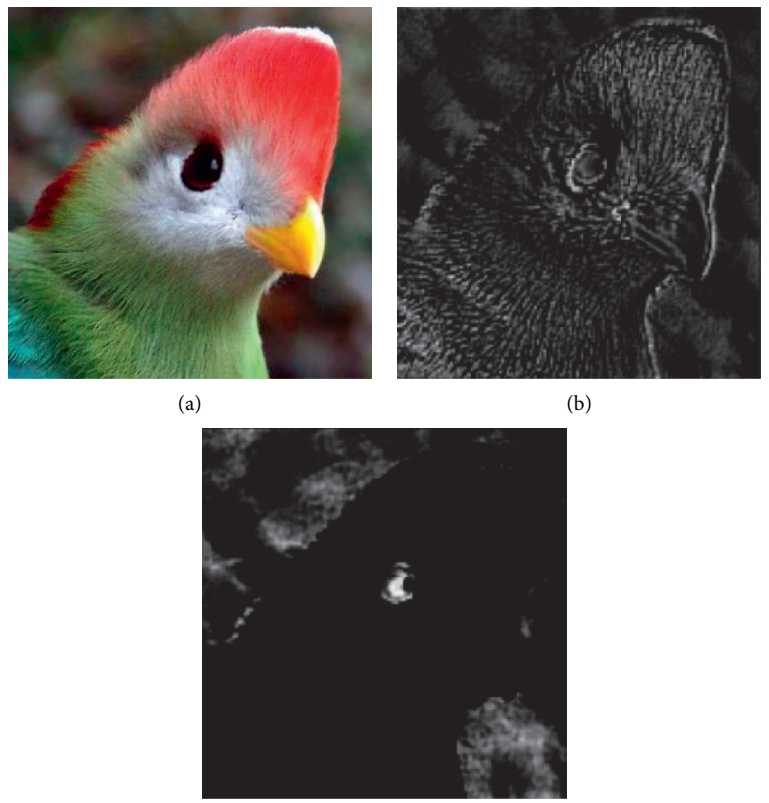
where p is sparse norm and λ is constant.

The implementation process of deconvolution is shown in Figure 3.

3.2.3. Scale Conversion. Scale problem is the core problem of target detection. In order to obtain feature maps with different resolutions with strong semantic information, we use the scale conversion method of feature map instead of the upsampling method in the original YOLOv3 network. Scale conversion is very efficient and can be directly embedded into dense blocks of Darknet. Assuming that the size of the input tensor of scale conversion is $H \times W \times (C \cdot r^2)$, where H and W are the length and width of the feature graph, $C \cdot r^2$ is the number of channels of the input feature graph, and r is the upsampling factor, this paper sets $r = 2$. Scale enlargement of feature map is shown in Figure 4.

It can be seen that reducing and enlarging the width and height of the transport layer is achieved by increasing and decreasing the number of channels, and the scale conversion module is an operation of periodic rearrangement of elements.

$$I_{x,y,c}^{SR} = I_{x/r,y/r,r\bmod(y,r)+\bmod(x,r)+cr^2}^{LR}. \tag{6}$$



(a) (b) (c)

Figure 1: Shallow features extracted by the convolution network. (a) Original picture; (b) texture information; (c) shape information.

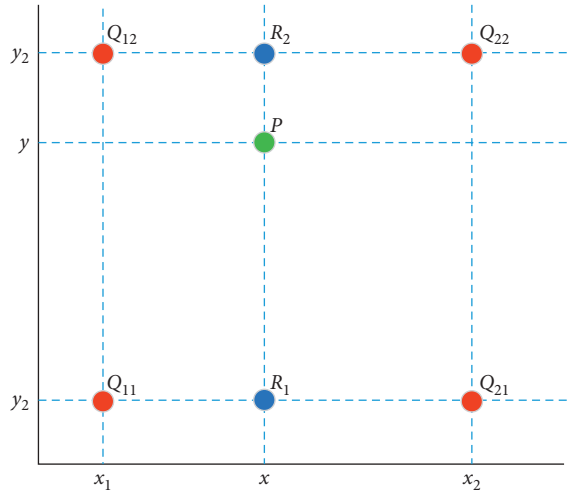


Figure 2: Schematic diagram of bilinear interpolation.

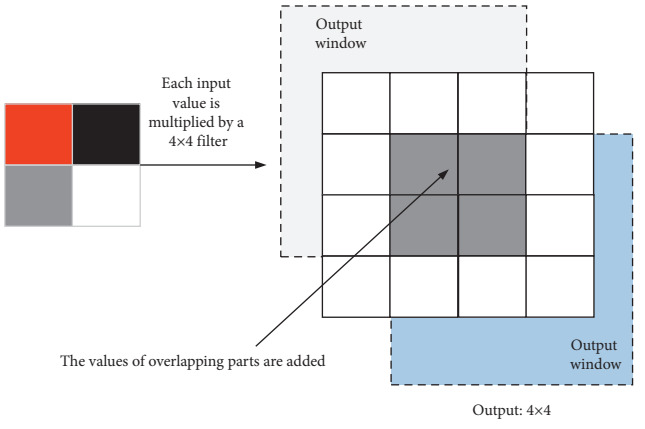


Figure 3: Implementation process of deconvolution.

where I^{SR} is a high-resolution feature map and I^{LR} is a low-resolution feature map. Scale transformation and use deconvolution layer must fill in zeros in the amplification step before convolution operation, without extra parameters and calculation overhead.

3.3. YOLOv3 Network Based on Feature Transformation, Migration, and Fusion. YOLO took the lead in innovatively combining the tasks of candidate selection stage and target recognition stage into one, and only one feature extraction can detect how many target objects and their positions [29]. Each grid in YOLO predicts two Bounding Box (BBox) in target detection. In YOLO's network structure, the task of extracting candidate regions is removed, and only the task of suggestion box loss regression exists. Therefore, the network structure is very simple, with only convolution and pooling operations, and finally, it is predicted by two fully connected layers, as shown in Figure 5.

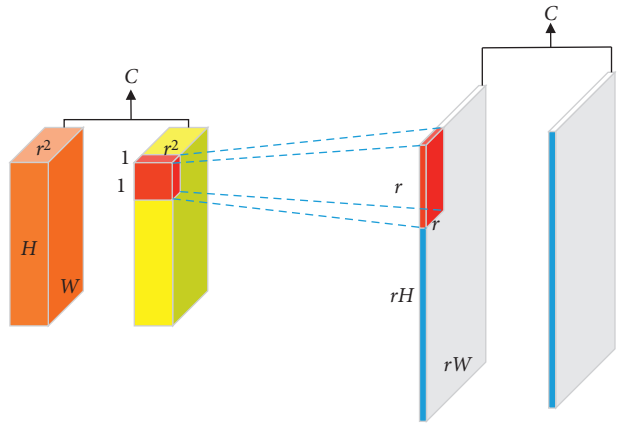


Figure 4: Scale enlargement of feature map.

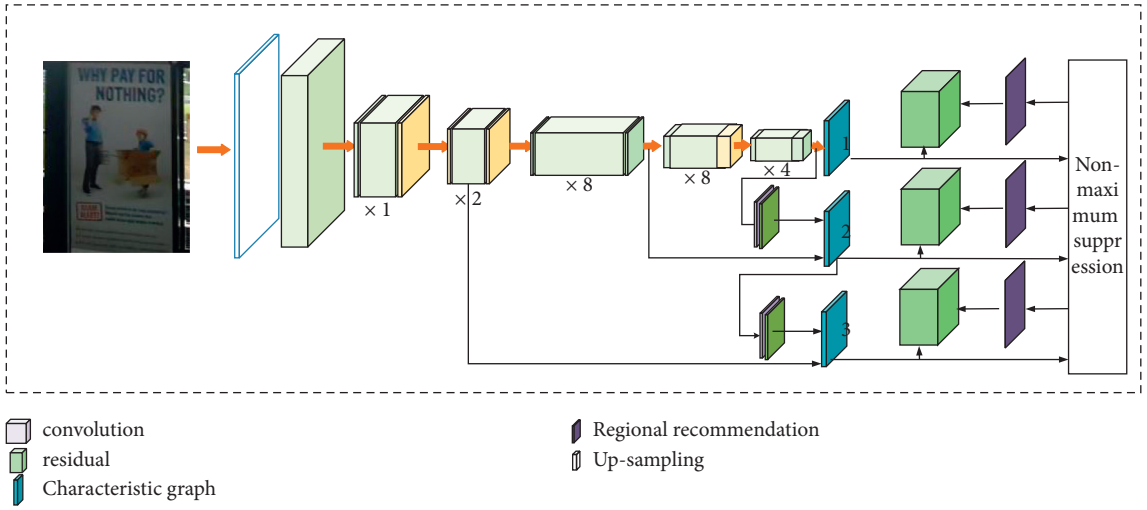


Figure 5: YOLO network structure.

It can be seen from Figure 5 that the first 24 convolution layers of the network are the backbone network for extracting image features, and finally, the extracted features are predicted through two fully connected layers. All-connection layer requires input, which is characterized by fixed dimensions. Each grid is predefined with 30-dimensional vector information corresponding to two suggestion boxes, as shown in Figure 6, which includes the positions of two BBox, the confidence of two BBox, and the classification probabilities of 20 objects.

During target detection, the specific category confidence score of each BBox is as follows:

$$P_{BBox} = P_r(\text{Class}_i) * IOU_{\text{pred}}^{\text{truth}}, \tag{7}$$

where $P_r(\text{Class}_i)$ is the probability of target occurrence and $IOU_{\text{pred}}^{\text{truth}}$ is the intersection over union (IOU) value of prediction frame and real frame.

In order to solve the gradient divergence problem caused by deepening the network model, YOLO3 borrowed the method of residual network and added the method of shortcut connections between some layers. The residual

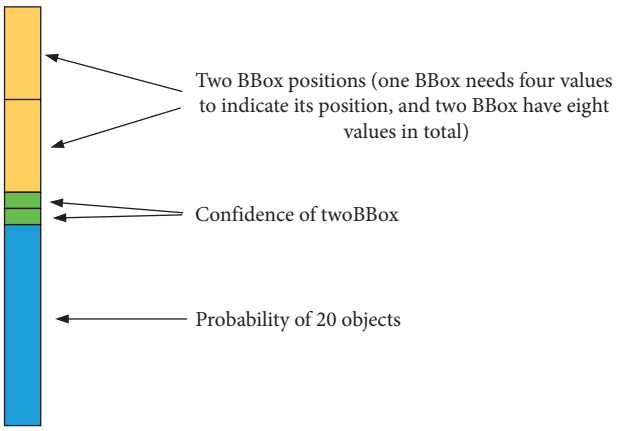


Figure 6: 30-dimensional output vector.

component of the specific direct connection method is shown in Figure 7 by directly transmitting the input x to the output, the output result is $f(x) + x$, and when $f(x) = 0$, then $H(x) = x$, the residual result approaches 0, and the training

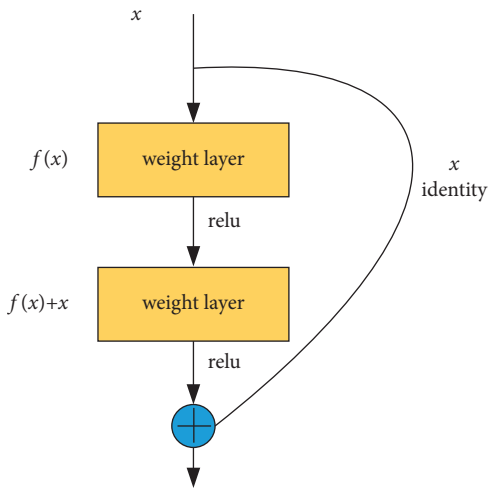


Figure 7: Residual components.

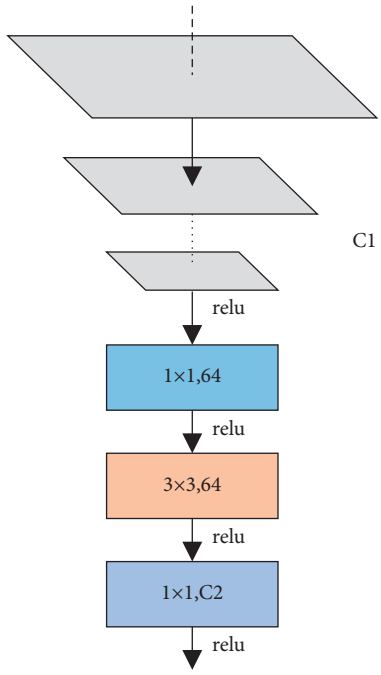


Figure 8: Scale-down migration fusion.

model converges, so that the accuracy will not decrease as the network deepens.

According to the above feature scale conversion operation, the scale size of the feature map is converted without destroying the original data of the feature. By using this feature scale transformation method, feature maps are reduced, migrated, and fused in the backbone network, and features are reused, so as to improve the expressive ability of features extracted by convolutional neural networks. The network layer with the same scale of input and output features is called the same level feature, and the last level feature of each level is selected as the reference feature because it has the strongest abstract expression ability after many convolution calculations. As shown in Figure 8, based on these reference features, this paper first reduces the feature scale of low-level features, sets the downsampling factor r to 2, performs convolution dimension reduction operation through 64 1×1 convolution kernels, then extracts features through convolution operation of 3×3 convolution kernels, then selects 1×1 convolution kernels matching the number of fusion layers to perform convolution dimension enhancement operation, and finally adds them to the fusion layers as the input of the subsequent network to continue extracting features. On the basis of the original YOLOv3 backbone network Darknet-53, this scale reduction migration fusion method is added to the feature layers with reference feature scales of 128×128, 64×64, 32×32, and 16×16.

At the same time, according to the abovementioned feature scale conversion operation, the feature scale amplification migration fusion method is adopted in the feature layers with reference feature scales of 8×8 and 16×16 in the trunk network Darknet-53, instead of the interpolation upsampling method which destroys the original data and has a huge amount of computation in the original FPN network. The specific implementation of feature scale amplification migration fusion is shown in Figure 9. First, the advanced feature map is subjected to feature scale amplification operation, the upsampling factor r is set to 2, and 64 1×1 convolution kernels are used for convolution dimension reduction operation; then, features are extracted by convolution operation of 3×3 convolution kernels, and then 1×1 convolution kernels matched with the number of fusion layers are selected for convolution dimension enhancement operation and finally added with the fusion layers as prediction features.

4. Experimental Results and Analysis

4.1. Experimental Environment and Data Set. The related software and hardware platforms are as follows: Intel Core i7 processor, 2.93 GHz computer with 8G memory, Ubuntu 16.04 LTS operating system, NVIDIA GPU with 24G memory, CUDA8.0, OPENCV3.2.0, and Darknet as the deep learning framework. The data set is ICDAR2015 data set, but the format of the data set needs to be changed into a trainable format accordingly, that is, image2voc, voc2 label files can convert data files into trainable data sets. Configure the internal classification standard of the model as a kind of TEXT, train according to the training set of the existing data set, and save the final model.

4.2. Determine Loss Function. As the loss function is the "baton" of the whole network learning and plays a very important role in the quality of the network model, it is necessary to design and optimize the loss function before network training. Therefore, during model training, the sum loss of square error is adopted for the coordinates, height and width of BBox, and the cross entropy loss is adopted for the classification scoring of BBox. The joint loss of multiple parts is as follows:

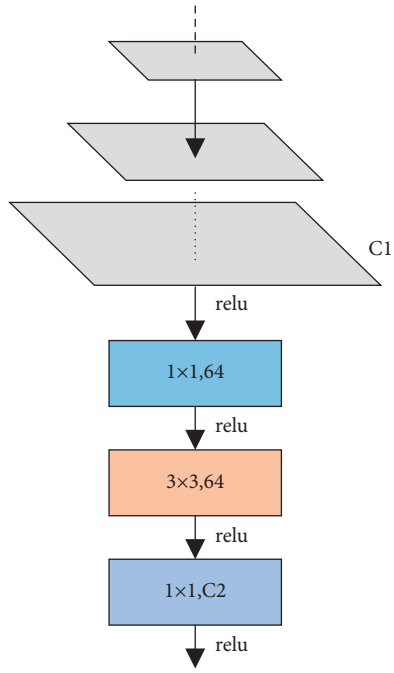


Figure 9: Scale-up migration fusion.

$$L = \sigma \sum_{i=1}^{N} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + \left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] + \lambda \sum_{i=1}^{N} \sum_{c \in C} -\hat{p}_i(c) \log\left(p_i(c)\right), \tag{8}$$

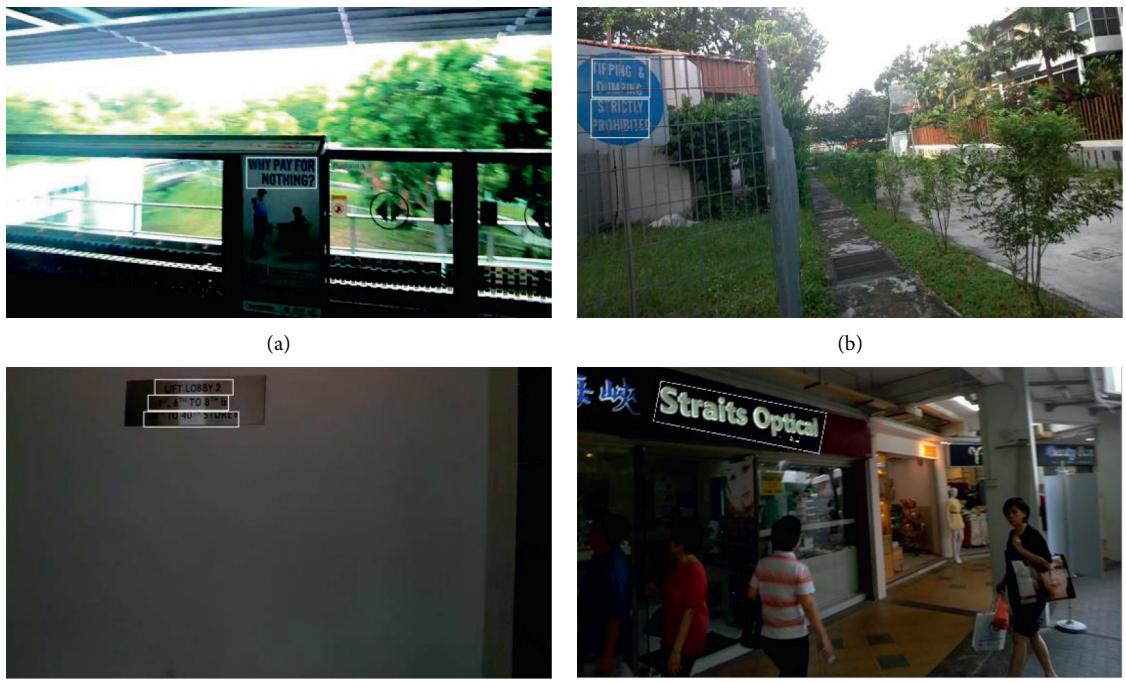
where N is the number of times when the IOU value of prior frame and real frame is greater than the threshold; (x_i, y_i), w_i and h_i are the center point coordinates, width and height of the i-th prediction frame; $(\hat{x}_i, \hat{y}_i)$; $\hat{w}_i$ and $\hat{h}_i$ are the center point coordinates, width and height of the real frame matched with the i-th prediction frame; $p_i(c)$ indicates the confidence score that the i-th prior frame belongs to category c; $\hat{p}_i(c)$ indicates the confidence score that the i-th prior frame matches the real frame belongs to category c; and σ and λ are the loss weights of BBox location and classification, respectively.

4.3. Set Network Parameters and Train. The training round of the proposed network on ICDAR2015 data set is 135. In the whole training process, the number of images in each batch is 64, the weight attenuation is 0.0005, and the momentum is 0.9. All additional network layer parameters are initialized with Xavier 8. The setting of learning rate is as follows: because the model tends to converge and be stable in the training and learning process, if the learning rate is kept high, the model will usually diverge due to unstable gradient. Therefore, the learning rate of the first 75 training rounds is set to 10^{-2}, the learning rate of the middle 30 training rounds is set to 10^{-3}, and the learning rate of the last 30 training rounds is set to 10^{-4}. After the network design is completed and the network parameters are set, the random gradient descent algorithm (SGD) is used to update the network parameters under the guidance of the loss function.

4.4. Result Analysis. In natural scenes, the text detection effect of traditional YOLOv3 network is not good, which may be due to the error in prediction when predicting the text pixels. In contrast, this text can be partially recognized in the improved YOLOv3 network, which improves the recall rate of locating targets. Text detection examples of traditional YOLOv3 network and improved YOLOv3 network are shown in Figures 10 and 11, respectively.

According to the effect comparison between Figures 10 and 11, the detection effect of the improved YOLOv3 network characters is better. Comparing the accuracy and detection speed of the proposed improved YOLOv3 network with the traditional YOLOv3 network, the results are shown in Table 1.

It can be seen from Table 1 that the MAP (mean average precision) of the proposed improved YOLOv3 network is increased by 9.5%, while the single frame detection time on a Tian X GPU is only increased to 27 ms from the original 22 ms. This shows that this chapter improves the algorithm of



(a) (b) (c) (d)

Figure 10: Text detection example of the traditional YOLOv3 network. (a) Subway advertisement. (b) Notice board. (c) Wall sign. (d) Signboard.

(a) (b) (c) (d)

Figure 11: Text detection example of the improved YOLOv3 network. (a) Subway advertisement. (b) Notice board. (c) Wall sign. (d) Signboard.

Table 1: Comparison of precision and detection speed

Network structure	mAP	Time (ms)
YOLOv3	67.2	**22**
Improved YOLOv3	**73.6**	27

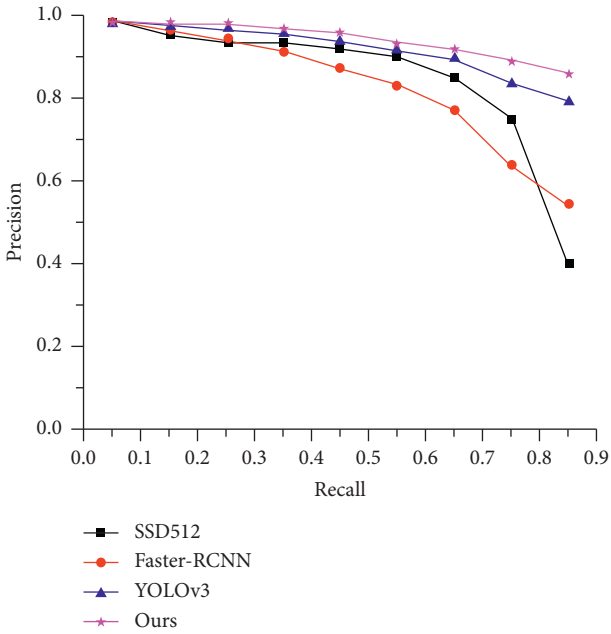


Figure 12: PR curve.

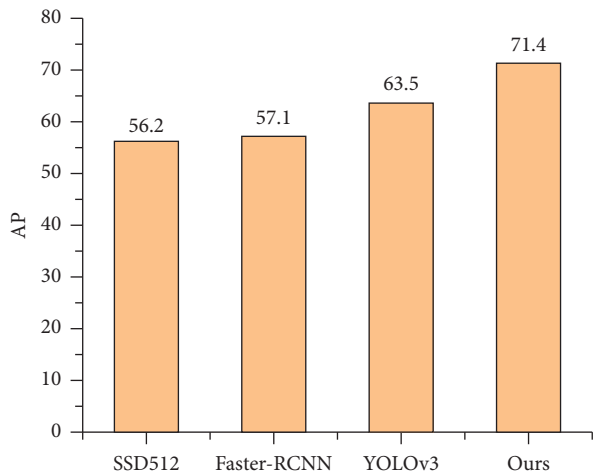


Figure 13: AP curve.

YOLOv3 network and does not bring too many parameters to slow down the detection speed of the model when constructing the backbone network of feature scale reduction migration fusion and FPN of feature scale enlargement migration fusion, thus improving the detection accuracy of the algorithm without affecting the excellent real-time detection performance of the original YOLOv3.

In addition, using AP (average precision) and PR (precision recall) as evaluation indexes, the performance of the improved YOLOv3 network is compared with that of the traditional YOLOv3 network, SSD512, and faster-RCNN. The PR curve and AP curve of each model on ICDAR2015 data set are shown in Figures 12 and 13, respectively.

5. Conclusions

In this paper, through the introduction of feature scale transformation method, a feature fusion method of feature scale enlargement migration fusion and feature scale reduction migration fusion is proposed. The original YOLOv3 network is modified by adopting the backbone network of feature scale reduction migration fusion to extract features and the FPN prediction of feature scale enlargement migration fusion. Compared with traditional YOLOv3, the improved YOLOv3 network through multiscale feature transformation, migration, and fusion provides more robust feature expression for the two tasks of object detection, frame regression, and category recognition. From the experimental results, it can be seen that the proposed method can effectively avoid the phenomenon of missing detection and wrong detection caused by occlusion or small target and obviously improve the detection accuracy of English text in natural scenes in videos.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. Demir and C. Bruzzone, "A novel active learning method in relevance feedback for content-based remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 539–556, 2015.

[2] L. Zhu, J. Shen, H. Jin, R. Zheng, and L. Xie, "Content-based visual landmark search via multimodal hypergraph learning," *IEEE Transactions on Cybernetics*, vol. 45, no. 12, pp. 2756–2769, 2015.

[3] S. R. Dubey, S. K. Singh, and R. K. Singh, "Multichannel decoded local binary patterns for content-based image retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4018–4032, 2016.

[4] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1216–1229, 2017.

[5] X. Wang, "Application of network protocol improvement and image content search in mathematical calculus 3D modeling video analysis[J]," *AEJ - Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4473–4482, 2021.

[6] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Information Sciences*, vol. 348, pp. 209–226, 2016.

[7] M. Hayat, S. H. Khan, M. Bennamoun, and S. An, "A spatial layout and scale invariant feature representation for indoor

scene classification," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4829–4841, 2016.

[8] D. A. Chandy, A. H. Christinal, A. J. Theodore, and S. E. Selvan, "Neighbourhood search feature selection method for content-based mammogram retrieval[J]," *Medical, & Biological Engineering & Computing*, vol. 55, no. 3, pp. 1–13, 2017.

[9] J. S. Deville, D. Kihara, and A. Sit, "2DKD: a toolkit for content-based local image search," *Source Code for Biology and Medicine*, vol. 15, no. 1, pp. 125–141, 2020.

[10] M. Moirangthem and T. R. Singh, "Brain tumor detection through content-based medical image retrieval using roi segmentation with Harmony search optimization," *Journal of Green Engineering*, vol. 10, no. 10, pp. 8939–8969, 2020.

[11] Z. Mehmood, M. Rashid, A. Rehman, T. Saba, H. Dawood, and H. Dawood, "Effect of complementary visual words versus complementary features on clustering for effective content-based image search," *Journal of Intelligent and Fuzzy Systems*, vol. 35, no. 5, pp. 5421–5434, 2018.

[12] M. Kurkure, A. Thakare, and S. Gudadhe, "Genetic candidate group search approach for post clustering content based image retrieval," *International Journal of Computer Application*, vol. 132, no. 16, pp. 6–9, 2015.

[13] B. Gábor, H. Szcs, and S. Dávid, "Content-based image retrieval for multiple objects search," *Cybernetics and Information Technologies*, vol. 17, no. 2, pp. 104–116, 2017.

[14] Z. Lan, T. Jung, K. Liu et al., "PIC: enable large-scale privacy preserving content-based image search on cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 11, pp. 3258–3271, 2017.

[15] M. E. Bates, "Picture this: challenges of video content and searchability," *Online*, vol. 41, no. 3, p. 72, 2017.

[16] X. Lin, X. Wang, and L. Li, "Intelligent detection of edge inconsistency for mechanical workpiece by machine vision with deep learning and variable geometry model," *Applied Intelligence*, vol. 50, no. 7, pp. 2105–2119, 2020.

[17] T. Kim, I. Y. Jung, and Y. C. Hu, "Automatic, location-privacy preserving dashcam video sharing using blockchain and deep learning," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–23, 2020.

[18] R. Minetto, N. Thome, M. Cord, N. J. Leite, and J. Stolfi, "SnooperText: a text detection system for automatic indexing of urban scenes," *Computer Vision and Image Understanding*, vol. 122, no. 5, pp. 92–104, 2014.

[19] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-orientation scene text detection with adaptive clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1930–1937, 2015.

[20] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8027–8048, 2014.

[21] Y.-Z. Zhuge and H.-C. Lu, "Robust video text detection with morphological filtering enhanced MSER," *Journal of Computer Science and Technology*, vol. 30, no. 2, pp. 353–363, 2015.

[22] S. M. Alqhtani, S. Luo, and B. Regan, "Fusing text and image for event detection in twitter," *The International journal of Multimedia & Its Applications*, vol. 7, no. 1, pp. 710–717, 2015.

[23] L. Tong, S. Palaiahnakote, C. L. Tan et al., "Video text detection systems," *Advances in Computer Vision & Pattern Recognition*, vol. 12, no. 4, pp. 379–388, 2014.

[24] R. Yang, X. Zha, K. Liu, and S. Xu, "A CNN model embedded with local feature knowledge and its application to time-varying signal classification," *Neural Networks*, vol. 142, no. 1, pp. 564–572, 2021.

[25] D. Wu, Q. Wu, X. Yin et al., "Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector," *Biosystems Engineering*, vol. 189, pp. 150–163, 2020.

[26] Y. Li, Z. Zhao, Y. Luo, and Z. Qiu, "Real-time pattern-recognition of GPR images with YOLO v3 implemented by tensorflow," *Sensors*, vol. 20, no. 22, p. 6476, 2020.

[27] F. Montalbo, "A computer-aided diagnosis of brain tumors using a fine-tuned YOLO-based model with transfer learning [J]," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 12, pp. 4816–4834, 2021.

[28] A. Shukla, I. Garkoti, A. M. B. Choudhary, and P. V. Dhaka, "Social distancing detection using open CV and yolo object detector[J]," *International Journal for Modern Trends in Science and Technology*, vol. 7, no. 1, pp. 93–95, 2021.

[29] L. Cao, H. Li, R. Xie, and J. Zhu, "A text detection algorithm for image of student exercises based on CTPN and enhanced YOLOv3," *IEEE Access*, vol. 8, pp. 176924–176934, 2020.