Journal of Healthcare Engineering

Computer Vision in Healthcare Applications

Lead Guest Editor: Yong Yang
Guest Editors: Dong S. Park, Pan Lin, and Junfeng Gao

Computer Vision in Healthcare Applications

Journal of Healthcare Engineering

Computer Vision in Healthcare Applications

Lead Guest Editor: Yong Yang
Guest Editors: Dong S. Park, Pan Lin, and Junfeng Gao

Editorial Board

Contents

Hindawi
Journal of Healthcare Engineering
Volume 2018, Article ID 5157020, 4 pages
https://doi.org/10.1155/2018/5157020

Editorial

Computer Vision in Healthcare Applications

Junfeng Gao,[1,2,3] **Yong Yang**,[4] **Pan Lin**,[1,2,3] **and Dong Sun Park**[5]

[1]*College of Biomedical Engineering, South-Central University for Nationalities, Wuhan 430074, China*
[2]*Key Laboratory of Cognitive Science, State Ethnic Affairs Commission, Wuhan 430074, China*
[3]*Hubei Key Laboratory of Medical Information Analysis and Tumor Diagnosis & Treatment, Wuhan 430074, China*
[4]*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, China*
[5]*IT Convergence Research Center, Chonbuk National University, Jeonju, Jeonbuk 54896, Republic of Korea*

Correspondence should be addressed to Yong Yang; yangyong5080@126.com

Received 27 December 2017; Accepted 28 December 2017; Published 4 March 2018

The research of computer vision, imaging processing and pattern recognition has made substantial progress during the past several decades. Also, medical imaging has attracted increasing attention in recent years due to its vital component in healthcare applications. Investigators have published a wealth of basic science and data documenting the progress and healthcare application on medical imaging. Since the development of these research fields has set the clinicians to advance from the bench to the bedside, the *Journal of Healthcare Engineering* set out to publish this special issue devoted to the topic of advanced computer vision methods for healthcare engineering, as well as review articles that will stimulate the continuing efforts to understand the problems usually encountered in this field. The result is a collection of fifteen outstanding articles submitted by investigators.

Following the goal of special issue, we identify four major domains covered by the papers. The first is medical image analysis for healthcare, the second is the computer vision for predictive analytics and therapy, the third is fundamental algorithms for medical images, and the last one focuses on the machine learning algorithms for medical images. Here, we give the review of these published papers.

1. Analysis of Medical Image

This theme attempts to address the improvement and new techniques on the analysis methods of medical image. First, integration of multimodal information carried out from different diagnostic imaging techniques is essential for a comprehensive characterization of the region under examination. Therefore, image coregistration has become crucial both for qualitative visual assessment and for quantitative multiparametric analysis in research applications. S. Monti et al. in Italy "An Evaluation of the Benefits of Simultaneous Acquisition on PET/MR Coregistration in Head/Neck Imaging" compare and assess the performance between the traditional coregistration methods applied to PET and MR acquired as single modalities and the obtained results with the implicitly coregistration of a hybrid PET/MR, in complex anatomical regions such as the head/neck (HN). The experimental results show that hybrid PET/MR provides a higher registration accuracy than the retrospectively coregistered images.

The feature extraction is one of the key issues for the analysis of medical images. I. I. Esener et al. in Turkey "A New Feature Ensemble with a Multistage Classification Scheme for Breast Cancer Diagnosis" develop a new and effective feature ensemble with a multistage classification which is used in a computer-aided diagnosis (CAD) system for breast cancer diagnosis. In this new method, four features, the local configuration pattern-based, statistical, and frequency domain features were concatenated as feature vectors, and eight well-known classifiers are used in a multistage classification scheme. High classification accuracy was obtained, and it shows that the proposed multistage classification scheme is more effective than the single-stage classification for breast cancer diagnosis.

Currently, the traditional approach to reduce colorectal cancer-related mortality is to perform regular screening in search for polyps, which results in polyp miss rate and inability to perform visual assessment of polyp malignancy. D. Vazquez et al. in Spain and Canada "A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images" propose an extended benchmark of colonoscopy image segmentation and establish a new strong benchmark for colonoscopy image analysis. By training a standard fully convolutional networks (FCN), they show that in endoluminal scene segmentation, the performance of FCN is better than the result of the prior researches.

2. Computer Vision for Predictive Analytics and Therapy

Computer vision technique has shown great application in surgery and therapy of some diseases. Recently, three-dimensional (3D) modeling and rapid prototyping technologies have driven the development of medical imaging modalities, such as CT and MRI. P. Gargiulo et al. in Iceland "New Directions in 3D Medical Modeling: 3D-Printing Anatomy and Functions in Neurosurgical Planning" combine CT and MRI images with DTI tractography and use image segmentation protocols to 3D model the skull base, tumor, and five eloquent fiber tracts. The authors provide a great potential therapy approach for advanced neurosurgical preparation.

The elderly is easy to fall and it will harm the body and accordingly has serious negative mental impacts on them. T.-H. Lin et al. in Taiwan "Fall Prevention Shoes Using Camera-Based Line-Laser Obstacle Detection System" design an interesting line-laser obstacle detection system to prevent the elderly from falls. In the system, a laser line passes through a horizontal plane and has a specific height to the ground, and optical axis in a camera has a specific inclined angle to the plane, and hence, the camera can observe the laser pattern to obtain the potential obstacles. Unfortunately, this system designed is useful mainly for indoor applications instead of outdoor environment.

Human activity recognition (HAR) is one of the widely studied computer vision problem. S. Zhang et al. in China "A Review on Human Activity Recognition Using Vision-Based Method" introduce an overview of various HAR approaches as well as their evolutions with the representative classical literatures. The authors highlight the advances of image representation approaches and classification methods in vision-based activity recognition. Representation approaches generally include global representations, local representations, and depth-based representations. They accordingly divide and describe the human activities into three levels including action primitives, actions/activities, and interactions. Also, they summarize the classification techniques in HAR application which include 7 types of method from the classic DTW and the newest deep learning. Lastly, they address that applying these current HAR approaches in real-world systems or applications has great challenge although up to now recent HAR approaches have achieved great success. Also, three future directions are recommended in their work.

3. Fundamental Algorithms for Medical Images

The majority of this issue focuses on the research of improved algorithm for medical images. Organ segmentation is a prerequisite for CAD systems. In fact, the segmentation algorithm is the most important and basic for image processing and also enhance the level of disease prediction and therapy. C. Pan et al. in China "Leukocyte Image Segmentation Using Novel Saliency Detection Based on Positive Feedback of Visual Perception" use the ensemble of polyharmonic extreme learning machine (EPELM) and positive feedback of perception to detect salient objects, which is totally data-driven without any prior knowledge and labeled samples compared with the existed algorithms. A positive feedback module based on EPELM focuses on fixation area for the purpose of intensifying objects, inhibiting noises, and promoting saturation in perception. Experiments on several standard image databases show that the novel algorithm outperforms the conventional saliency detection algorithms and also segments nucleated cells successfully in different imaging conditions.

High-intensity focused ultrasound (HIFU) has been proposed for the safe ablation of both malignant and benign tissues and as an agent for drug delivery, while MRI has been proposed for guidance and monitoring for the therapy. A. Vargas-Olivares et al. in México and Canada "Segmentation Method for Magnetic Resonance-guided High-Intensity Focused Ultrasound Therapy Planning" used the MR images for the HIFU therapy planning and propose an efficient segmentation approach. The segmentation scheme uses the watershed method to identify the regions found on the HIFU treatment. In addition, the authors also propose a thread pool strategy, in order to reduce the computational overload of the processing time of the group of MR images and the segmentation algorithm.

Recently, random walkers (RW) have attracted a growing interest to process segmentation of medical images. However, classical RW method needs a long computation time and a high memory usage because of the construction of corresponding large-scale graph to solve the resulting sparse linear system. C. Dong et al. in China and USA "An Improved Random Walker with Bayes Model For Volumetric Medical Image Segmentation" incorporate the prior (shape and intensity) knowledge in the optimization of sparse linear system. Integrating the Bayes model into the RW sparse system, the organ is automatically segmented for the adjacent slice, which is called RWBayes algorithm in the article. Compared with the conventional RW and the state-of-the-art interactive segmentation methods, their method can significantly improve the segmentation accuracy and could be extended to segment other organs in the future.

Automatic segmentation of the spinal cord in MR images remains a difficult task. C.-C. Liao et al. in Taiwan "Atlas-Free Cervical Spinal Cord Segmentation on Midsagittal T2-Weighted Magnetic Resonance Images" present an automatic segmentation method on sagittal T2-weighted images.

The method is atlas-free, in which expectation maximization algorithm is used to cluster the pixels on a midsagittal MR image according to their gray levels or SIs. Dynamic programming is used to detect anatomical structures and their edges. The detection of the anterior and posterior edges of the spinal cord within the cervical spinal canal is finally successful in all 79 images, showing its high accuracy and robustness. Based on this proposed algorithm, using alone or combining with others, one can develop a computer-aided diagnosis system with massive screening on cervical spine diseases. Finally, the authors point out several limitations in the algorithm, such as its inability to be applied to lower lumbar spinal levels.

The misalignments originated from motion and deformation often result in errors in estimating an apparent diffusion coefficient (ADC) map fitted with prostate DWI, and the ADC map is an important indicator in diagnosing prostate cancer. Until now, there are few studies that focus on this misalignment in prostate DWI. L. Hao et al. in China "Nonrigid Registration of Prostate Diffusion-Weighted MR" apply affine transformation to DWI to correct intraslice motions. Then, nonrigid registration based on free-form deformation (FFD) is used to compensate for intraimage deformations. The experimental results show that the proposed algorithm can correct the misalignment of prostate DWI and decrease the artifacts of ROI in the ADC maps. These ADC maps thus obtain sharper contours of lesions, which are helpful for improving the diagnosis and clinical staging of prostate cancer.

Medical ultrasound is widely used in the diagnosis and assessment of internal body structures and also plays a key role in treating various diseases due to its safety, noninvasion, and well tolerance in patients. However, the images are always contaminated with speckle noise and hence hinder the identification of image details. Currently, many methods have been proposed to remove the noise and preserve the image details at the same time. M. Szczepański and K. Radlak in Poland "Digital Path Approach Despeckle Filter for Ultrasound Imaging and Video" propose a so-called escaping paths based on traditional digital paths, and also, they extend this concept from the spatial domain (2D) to the spatiotemporal domain (3D) that is designed for multiplicative noise suppression, specifically for ultrasound image and video filtering. In addition, the extended neighborhood model is used to increase the filter denoising ability, which is based on von Neumann concept derived from cellular automata theory. The experimental results prove that the proposed removal technique outperforms the state-of-the-art approach for multiplicative noise removal with lower computational overload which enables one to complete image processing tasks and image enhancement of video streams in a real-time environment.

A primary challenge in accelerating MR imaging is how to reconstruct high-resolution images from undersampled k-space data. There is a trade-off between the spatial resolution and temporal resolution. J. Chen et al. in China "Low-Rank and Sparse Decomposition Model for Accelerating Dynamic MRI Reconstruction" introduce a low-rank and sparse decomposition model to resolve this problem, which is based on the theory of robust principal component analysis (RPCA). Unlike k-t RPCA (a method that uses the low-rank plus sparse decomposition prior to reconstruction of dynamic MRI from part of the k-space measurements), the authors propose inexact augmented Lagrangian method (IALM) to solve the optimization of RPCA and to accelerate the dynamic MRI reconstruction from highly undersampled k-space data, which has a generalized formulation capability of separating dynamic MR data into low-rank and sparse component. The experimental results on cardiac datasets prove that the proposed method can achieve more satisfactory reconstruction performance and faster reconstruction speed, compared with the state-of the-art reconstruction methods.

4. Machine Learning Algorithms for Medical Images

The growth of the older adult population in the world is surprising and it will have a great impact on the healthcare system. The elders always lack self-care ability and hence, healthcare and nursing robot draw much attention in recent years. Although somatosensory technology has been introduced into the activity recognition and healthcare interaction of the elderly, traditional detection method is always in a single modal. In order to develop an efficient and convenient interaction assistant system for nurses and patients with dementia, X. Dang et al. in China "An Interactive Care System Based on a Depth Image and EEG for Aged Patients with Dementia" propose two novel multimodal sparse autoencoder frameworks based on motion and mental features. First, the motion is extracted after the preprocessing of depth image and then EEG signals as the mental feature is recorded. The proposed novel system is designed to be based on the multimodal deep neural networks for the patient with dementia with special needs. The input features of the networks include (1) extracted motion features based on the depth image sensor and (2) EEG features. The output layer is the type recognition of the patient's help requirement. Experimental results show that the proposed algorithm simplifies the process of the recognition and achieved 96.5% and 96.4% (accuracy and recall rate), respectively, for the shuffled dataset, and 90.9% and 92.6%, respectively, for the continuous dataset. Also, the proposed algorithms simplify the acquisition and data processing under high action recognition ratio compared with the traditional method.

Recently, deep learning has become very popular in artificial intelligence. Q. Song et al. in China "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images" employ a convolution neural network (CNN), deep neural network (DNN), and stacked autoencoder (SAE) for the early diagnosis of lung cancer to doctors. The experimental results suggest that the CNN archived the best performance than DNN and SAE.

N. D. Kamarudin et al. in Malaysia and Japan "A Fast SVM-Based Tongue's Colour Classification Aided by k-Means Clustering Identifiers and Colour Attributes as Computer-Assisted Tool for Tongue Diagnosis" propose a two-stage classification system for tongue color diagnosis aided with the devised clustering identifiers, and it can diagnose three tongue colors: red, light red, and deep red. The diagnosis system is very useful for the early detection of imbalance condition inside the body. Experimental result shows that this novel classification system outperforms the conventional SVM by 20% computational time and 15% in terms of classification accuracy.

5. Conclusion

These authors highlight both the promise and the challenges faced by this healthcare application field of medical images. Their researches identify the critical need for clinical and theory prospective of medical images. This special issue brings about various new developments in computer vision about medical images and clinical application. In summary, this special issue provides a snapshot of the computer vision in healthcare applications on medical images across the globe. Hopefully, this publication will provide a good reference for future computer vision, analysis algorithms, and machine learning of medical images. However, there are still some key messages that emerge from the papers compiled within this special issue: there still remain limitation and challenge for computer vision and various algorithms and processing techniques of medical images although these works show good efficiency than traditional and state-of-art methods. We hope that this theme issue will further advance our understanding of computer vision about medical image processing and healthcare applications and pave the way for new directions in medical images and computer vision research across health and disease.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (61773408, 81271659, 61662026, and 61473221). The guest editors are very thankful to all the anonymous reviewers of the journal and the perseverant and generous support of the editor in chief.

Junfeng Gao
Yong Yang
Pan Lin
Dong Sun Park

Hindawi
Journal of Healthcare Engineering
Volume 2018, Article ID 5098973, 11 pages
https://doi.org/10.1155/2018/5098973

Research Article

Leukocyte Image Segmentation Using Novel Saliency Detection Based on Positive Feedback of Visual Perception

Chen Pan,[1] **Wenlong Xu**,[1] **Dan Shen**,[2] **and Yong Yang**[3]

[1]*China Jiliang University, Hangzhou, Zhejiang 310018, China*
[2]*Department of Hematology at the First Hospital Affiliated to Medical College, Zhejiang University, Hangzhou, Zhejiang 310003, China*
[3]*School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China*

Correspondence should be addressed to Chen Pan; pc916@cjlu.edu.cn

Received 24 February 2017; Revised 8 November 2017; Accepted 21 November 2017; Published 1 February 2018

Academic Editor: Maria Lindén

This paper presents a novel method for salient object detection in nature image by simulating microsaccades in fixational eye movements. Due to a nucleated cell usually stained that is salient obviously, the proposed method is suitable to segment nucleated cell. Firstly, the existing fixation prediction method is utilized to produce an initial fixation area. Followed EPELM (ensemble of polyharmonic extreme learning machine) is trained on-line by the pixels sampling from the fixation and nonfixation area. Then the model of EPELM could be used to classify image pixels to form new binary fixation area. Depending upon the updated fixation area, the procedure of "pixel sampling-learning-classification" could be performed iteratively. If the previous binary fixation area and the latter one were similar enough in iteration, it indicates that the perception is saturated and the loop should be terminated. The binary output in iteration could be regarded as a kind of visual stimulation. So the multiple outputs of visual stimuli can be accumulated to form a new saliency map. Experiments on three image databases show the validity of our method. It can segment nucleated cells successfully in different imaging conditions.

1. Introduction

Microscopic leukocyte analysis is a powerful diagnostic tool for many types of diseases for which it is vital to recognize and count different lineages and maturity levels of leukocytes. Computer-aided automatic analysis not only saves manpower and time but also reduces human error. The most important step in automatic image analysis is segmentation. Human leukocytes (WBCs) are colorless. Blood and bone marrow smears are conventionally prepared with Wright-Giemsa stain in order to visualize obviously and identify WBCs. However, different smear preparation and imaging conditions may result in large biases and changes in image color. It is difficult to segment entire leukocyte populations since color distributions may be uncertain.

Nature image is typical unstructured data. Modeling such data via machine learning has been a hotspot for decades. In recent years, two classes' learning-based algorithms, bottom-up and top-down, which composed with shallow and deep neural networks, respectively, are widely used to solve the segmentation problem. In bottom-up framework, literatures [1, 2] had proposed data-driven methods to segment leukocyte image via "pixel sampling-learning-classification" procedure based on shallow network (SVM or ELM). However, those algorithms have some priori restrictions. For example, the algorithm assumes that the nuclei of WBCs are surrounded by cytoplasm and are always deeply stained so that the object intensity is low while the background intensity is bright. Moreover, preparing the training samples is very critical in those methods. Because if the samples are not good (or not pure) enough, the learning-based algorithm may output undesired object. It is a challenge to solve such noise-sample-sensitive problem in a learning-based framework. The state-of-the-art top-down model is developed from deep learning, which has been successfully used for image segmentation [3]. Deep learning-based algorithm reflects the best performance in many applications so far, since that can deal with object-level or image-level representative features from training samples. However, deep networks often have huge parameters than the shallow one, so that they need massive labeled sample data to tune parameters repeatedly in training. The most of the existing top-down learning-based methods are time-consuming in off-line training process or labeling positive samples manually.

Due to nucleated cell usually stained with salient color, Zheng et al. [4] firstly locate nuclei by saliency detection method and then extract the nucleated cell by maker-controlled watershed. However, saliency detection in Zheng's method was too simple and may be out of date. It is necessary to find some new ways to update them.

In general, without prior knowledge and effective samples, many segmentation methods may fail in practice. In addition, we note that the information is often feedforward and lacks feedback process in most bottom-up or top-down models. It greatly differs from human vision. That may be one of the reasons the performance of machine vision is far from that of human vision.

Human accepts attention by making a series of eye movements. There are two forms of eye movement: saccades and microsaccades. (1) In saccade stage, human eyes aim to find candidate object so it makes sharply shifts in the whole field of view. (2) While candidates are identified as target, the eyes will make a series of dense tiny movements that is called microsaccades around the target for the purpose of intensify objects and inhibit noises. Continuous microsaccades will lead to visual fading [5], and the eye movement will switch to the stage of saccades to find new objects. The integration of saccades and microsaccades contributes to the quick and efficient performance of human vision system.

Motivated by the above reasons, this paper presents a novel saliency detection framework by simulating microsaccades and visual fading, without prior knowledge and labeled samples. We construct a positive feedback loop to focus on fixation area and intensify objects repeatedly. Ensemble of polyharmonic extreme learning machine (EPELM) [6, 7] is utilized to simulate the human neural system to produce visual stimulus. Depending on sampling from previous fixation area (input), training EPELM model using the samples, and classifying image pixels by EPELM, new fixation area can be output in iteration. If the input and output fixation area were similar enough, it indicates that the perception is saturated and the iteration should be terminated. The final fixation area is the segmentation result in our method. Experimental results show that new saliency method with positive feedback loop can achieve better performance and can greatly improve the performance of the existing saliency detection methods.

In summary, the main contributions of our work are as follows. (1) We propose a novel learning-based algorithm to detect salient objects depending on bottom-up saliency. It is good to segment stained leukocyte without any prior knowledge and labeled samples. (2) A positive feedback module based on EPELM is presented which focuses on fixation area for the purpose of intensifying objects, inhibiting noises, and promoting saturation in perception. Positive feedback of perception may be indispensable in saliency detection.

In the rest of the paper, we introduce the works of saliency detection that are heavily related to our approach in Section 2. Then we describe our algorithm in Section 3 and finally discuss the experimental result in Section 4. Section 5 is the conclusions.

2. Related Works

Visual attention is a remarkable capability of early primate visual system, which helps human complete scene analysis in real-time with limited resources. Inspired by it, various computational models, which are called saliency models, have been proposed according to the psychological and neurobiological findings. Saliency model aims to identify the most salient foreground object from the background, and this problem in its essence is a figure/ground segmentation problem. In general, saliency detection can also be grouped into two categories: top-down methods and bottom-up methods. Bottom-up methods are rapid, data-driven, and task-independent, which construct saliency maps based on low-level visual information, such as pixel level or super pixel level. Due to the absence of high-level knowledge, all bottom-up methods rely on assumptions about the properties of objects and backgrounds. The widely utilized assumptions could be contrast prior, boundary prior, center prior, background prior, and so on. In contrast, top-down approaches are slower, volition-controlled, and task-driven and require supervised learning based on training samples with manual labels.

The classic bottom-up computational model is Borji et al.'s method [8], which gets saliency values of each pixel by center-surround contrast. Hou and Zhang [9] use the residual Fourier amplitude spectrum to form saliency map. Both of the above two models aim to predict human fixation points; hence, saliency maps computed by these models are spatially discontinuous. While at the same time, models for the purpose of salient region detection have been proposed in [10], Goferman et al. proposed a saliency model based on context-aware, and Cheng et al. [11] presented global contrast-based saliency computation methods, called histogram-based contrast (HC) and spatial information-enhanced region-based contrast (RC). These types of model can generate saliency maps with fine details and high resolution. Literature [8] indicates that models for salient region detection shown actual advantage in contrast with models for fixation prediction in terms of various computer vision applications.

Recently, many learning models are proposed for saliency computation. Methods based on supervised learning have emerged [12], and these approaches use large fine annotation images to train saliency model, which is a typically knowledge-driven approach. At the same time, there are some unsupervised learning approaches [13]. All of the above methods either rely on large manual-labeled dataset and off-line training or lie on numbers of parameters set for modeling.

In the bottom-up framework, [14–16] presented some effective ways to to train a set of weak classifiers based on initial saliency maps and then obtained a strong model by integrating the weak classifiers or their results. In their approaches, multiple types of classifiers, multiscale analysis, and graph cut algorithm could be taken together, such as Na et al. [14] presented BL algorithm which detect objects via bootstrap learning of SVMs. Huang et al. [15] presented MIL algorithm depending on object proposals and multiple instance learning using SVMs. And in [16], Zhang et al. presented salient object detection based on ELM. Those methods are very attractive to us since the learning-based approach is similar to human brain, and their results are very close to human perception. However, we noticed that those methods are almost no reference to human visual mechanism. They implement function from computing rather than simulating human vision. It is time-consuming because of multiscale analysis and multimodel parallel computing that may lose the speed of bottom-up model. In addition, the information is also feedforward and lacks feedback process in those methods. There may be large room to improve them.

3. Method

Our method consists of four main components:

(i) Generating a gaze area

(ii) Forming a preliminary object using coarse saliency map by learning

(iii) Suppressing the background

(iv) Intensifying object to form a sense of saturation by learning-based positive feedback

The framework of our method is illustrated in Figure 1.

The key steps are listed as follows. See Figure 2, the main pipeline of our method.

Step 1. An initial saliency map is made from input image by SR algorithm.

Step 2. Use coarse saliency detection by EPELM learning:

(1) Sort pixels according to the saliency, and select the first n pixels with large value ($n = 100$ in our experiment).

(2) The selected pixels form a minimum rectangle box containing them. Inside the box is the fixation area, so the outside is the nonfixation area.

(3) Random sample m pixels with high gradient are from the fixation area (positive samples). And random sample equal pixels are from the nonfixation area (negative samples) ($m = 500$ in our experiment).

(4) Use training EPELM using the positive and negative pixels with RGB features.

(5) Classify image pixels by EPELM. Each binary output of PELM is regarded as single stimulation, could be normalized, and is added to form a coarse saliency map.

Step 3. RBD algorithm is used to reduce the noise in the coarse saliency map, by background detection and saliency optimization.

Step 4. Intensify objects using positive feedback loop:

(1) Threshold the optimized saliency map to make new binary fixation area (BW_i).

(2) If BW_i-1 has been existed, then judge whether BW_i is similar enough to BW_i-1. If true, go to step 5 (break the loop); else, do the next step.

(3) Use Saliency detection by EPELM learning (same as step 2). Each binary output of PELM could be normalized and added to the saliency map.

(4) Return to step 1 in the current step.

Step 5. The final segmentation result is BW_i (end).

3.1. The Function of SR and RBD Algorithms. SR (spectrum residual) method was presented by Hou and Zhang [9], which aims to predict human fixations and often produces blob-like and sparse saliency map corresponding to the human fixation spots on scenes. Let $I(x)$ be the image, x be the pixel position, $F()$ be the Fourier transformation; then

$$\begin{aligned} A(f) &= |F([I(x)])|, \\ P(f) &= \varphi(F[I(x)]), \\ L(f) &= \log(A(f)), \\ R(f) &= L(f) - h_n(f) * L(f), \\ \mathrm{SR}(x) &= \left|F^{-1}[\exp\{R(f) + jP(f)\}]\right|^2, \end{aligned} \tag{1}$$

where $A(f)$ is the amplitude spectrum of image, $P(f)$ is the phase spectrum of image, $L(f)$ is the log of amplitude spectrum, $R(f)$ represents residual Fourier amplitude spectrum, $\mathrm{SR}(x)$ is the saliency map, $\varphi()$ is the operation to extract phase, and $h_n(f)$ is an average operator.

The salient points detected by SR often have strong correlation with eye gaze spots. Besides, SR is very similar to human perception since saliency map may change when the scale of the image changes. And it is one of the fastest fixation prediction algorithms [8]. So we select it to simulate human fixation.

In our method, we firstly provide an initial fixation area using SR, then sampling from there, and learning by EPELM. Multiple random sampling may be equivalent to the micro scan in the fixation region. Because the training samples are few ($m = 500$ in this paper), the EPELM classifier can be trained in real-time. After that, those models are used to classify image pixels into classes of object or background. The binary output of every PELM model could be

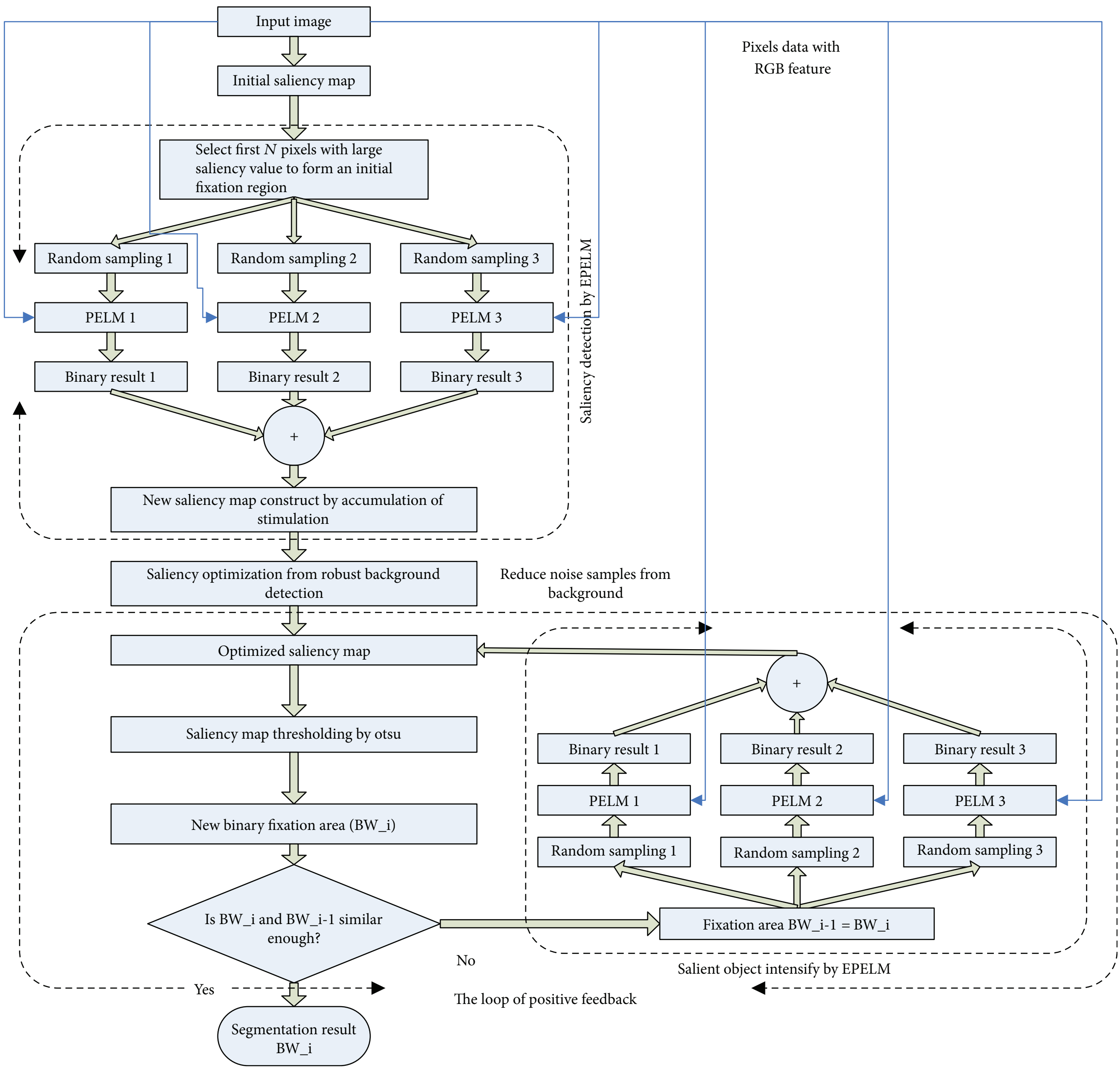


Figure 1: The framework of the proposed method.

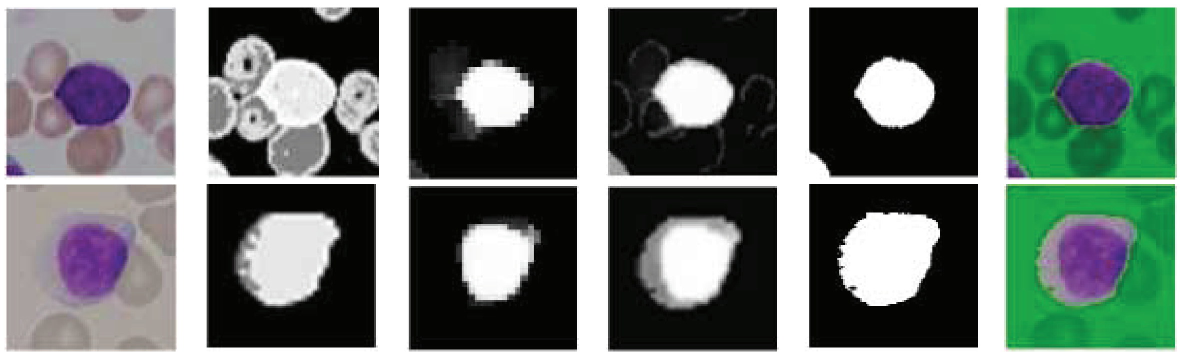
treated as a kind of stimulus just like neuron firing in human brain. Multiple outputs of PELMs could be accumulated together and normalized to form a new coarse saliency map. Figures 3(a)–3(f) show an example.

Since the initial fixation area is often rough (see the red box in Figure 3(a)), so that there is a lot of noise in the positive and negative samples. Such noise samples may easily lead to undesired output. Although we accumulate the learning-based results, it is not enough to decrease the bad influence of background pixels to foreground. In order to erase the error caused by noise samples, the coarse saliency map needs to be optimized further by suppressing the background.

RBD (saliency optimization from robust background detection) was proposed by Zhu et al. [17], which belongs to salient object detection models and attempts to highlight the whole salient object by suppressing the background. Zhu et al. proposed a robust background measure, called boundary connectivity. It characterizes the spatial layout of image regions with respect to image boundaries. It is defined as

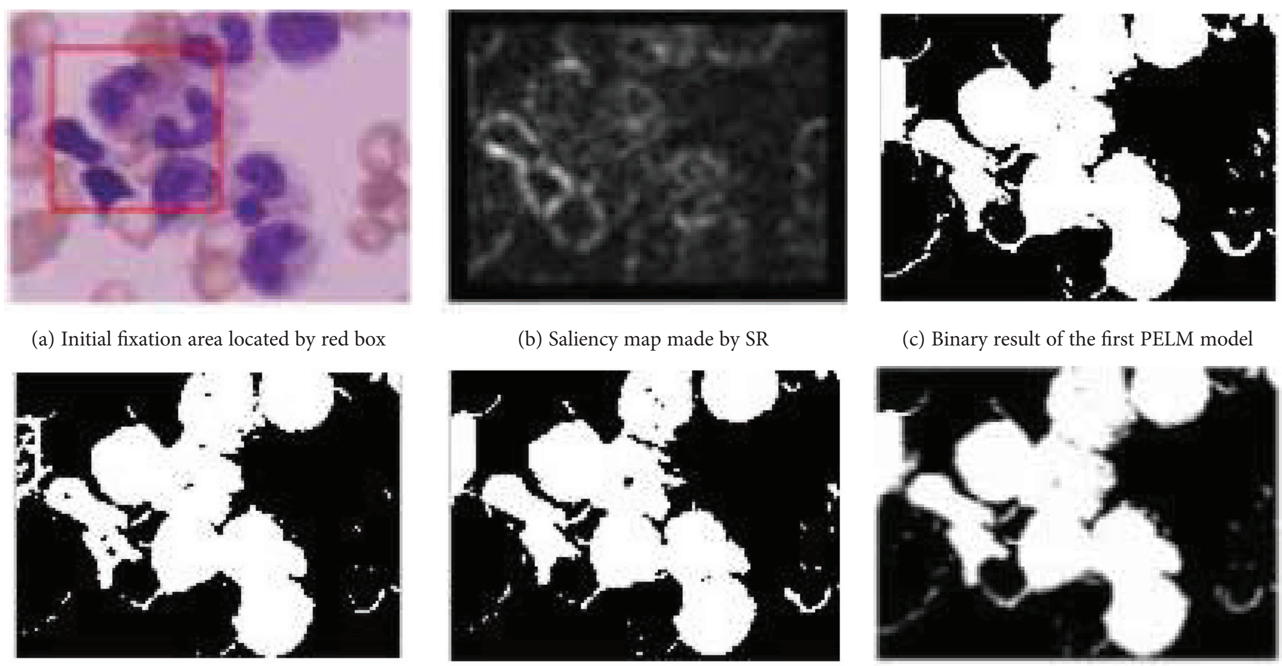
$$\text{BndCon} = \frac{|\{p|p \in R, p \in \text{Bnd}\}|}{\sqrt{|\{p|p \in R\}|}}, \quad (2)$$

where p is an image patch and Bnd is the set of image boundary patches. It has an intuitive geometrical interpretation: it is



Input image (a) | The coarse saliency detection by EPELM (b) | Optimized saliency map by RBD (c) | Saliency map after positive feedback (d) | Final fixation area (e) | Color object (f)

FIGURE 2: The main pipeline of our method.



(a) Initial fixation area located by red box

(b) Saliency map made by SR

(c) Binary result of the first PELM model

(d) Binary result of the second PELM model

(e) Binary result of the third PELM model

(f) The coarse saliency map by accumulating (c), (d), and (e)

FIGURE 3: (a-b) Initial fixation area made by SR; (c–f) the coarse saliency map made by PELM learning.

the ratio of a region's perimeter on the boundary to the region's overall perimeter or square root of its area.

Zhu et al. presented an approach depending on superpixels to compute background probability by boundary connectivity.

$$\omega_i^{\mathrm{bg}} = 1 - \exp\left(\frac{\mathrm{BndCon}^2(p_i)}{2\sigma_{\mathrm{BndCon}}^2}\right). \quad (3)$$

The salient object detection problem in their model is regarded as the optimization of the saliency values of all image superpixels. An optimization framework to integrate an initial saliency map with the background measure is presented. The objective cost function is designed to assign the object region value 1 and the background region value 0, respectively. The optimal saliency map is then obtained by minimizing the cost function.

Let the saliency values of N superpixels be $\{s_i\}_{i=1}^N$, and the cost function is

$$\mathrm{cost} = \sum_{i=1}^{N} \omega_i^{\mathrm{bg}} s_i^2 + \sum_{i=1}^{N} \omega_i^{\mathrm{fg}} (s_i - 1)^2 + \sum_{ij} \omega_{ij} (s_i - s_j)^2. \quad (4)$$

There are three terms which define costs from different constraints. ω_i^{bg} is the background probability, ω_i^{fg} is the foreground probability often represented by initial saliency map, and ω_{ij} is the smoothness term which encourages continuous saliency values which is used to erase small noise in both

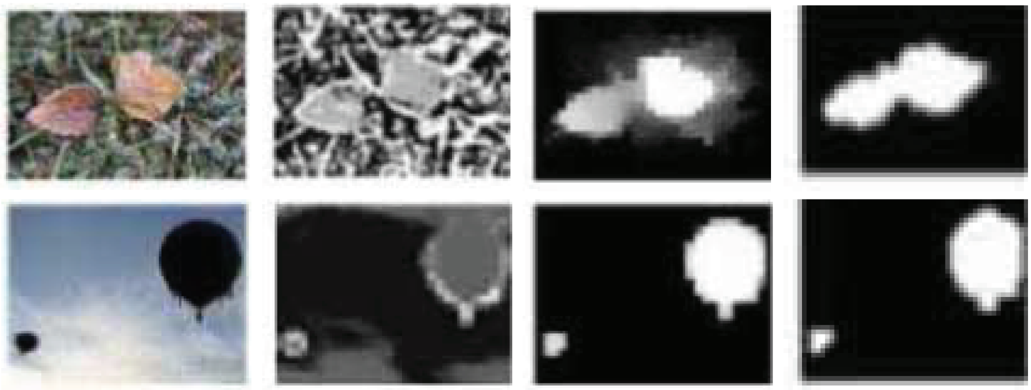


Figure 4: The function of RBD algorithm: suppressing the background pixels.

background and foreground terms. We select optimal value of $\{s_i\}_{i=1}^{N}$ to minimize the cost.

In our method, ω_i^{fg} is the coarse saliency map produced by our method from step 1 to step 2. It could be carried into RBD algorithm to eliminate noise effects of background. See Figure 4. Pure samples will yield more precision model and binary result through the followed positive feedback.

3.2. Training of EPELM. ELM (extreme learning machine) has been widely used as a fast learning method for feedforward networks with a single-hidden layer [6]. Recently, Zhao et al. [7] extended it for more stable performance, which is called EPELM (ensemble of polyharmonic extreme learning machine). It has shown good performance in human face recognition. Due to polyharmonic mechanism, EPELM is an effective way to deal both kinds of scattered date with rapid changed and slow variations. Different from traditional learning algorithms which based on the gradient descent techniques for parameter optimization, EPELM sets its inner weights randomly and needs no iterative training. It can be trained on-line with small sample sets and needs not tuned with any parameter. So we use EPELM for learning-based saliency detection.

For a given set of training samples $\{(x_i, t_i)\}_{i=1}^{N} \subset R_n \times R_m$, the output of a PELM with L hidden nodes can be written by

$$f^r(x) = \sum_{i=1}^{L} \beta^r{}_i \cdot G\left(a_i{}^{(r)}, b_i{}^{(r)}, x\right) + P(x), x \in R^n, \quad (5)$$

where a_i and b_i are the inner weights of input node to hidden node. β_i is the output weights of hidden node to output node. The inner weights in this model are randomly assigned. $G(a_i, b_i, x)$ is the output of ith hidden note. $p(x)$ is a polynomial with low degree, which can deal with the type of data with slow variations. Output weights β can be computed by the following formula:

$$\hat{\beta}^r = H^+ T, \quad (6)$$

where H^+ is the Moore-Penrose pseudoinverse of the hidden layer output matrix, and $T = [t_1, t_2, \dots, t_n]^T$.

For the aim of gaining more stable model, we integrate numbers of PELM. The parameter p denotes the number of PELM grouped in the EPELM. The function for EPELM is ($p = 3$ in our experiment)

$$f(x) = \frac{1}{p}\sum_{r=1}^{p} f^r(x). \quad (7)$$

In this paper, EPELM can be treated as neural system of human brain to accept stimulus and output new one. The function of positive feedback loop based on EPELM is illustrated in Figure 5. It is easy and quick that visual perception becomes saturated in positive feedback loop.

4. Experimental Results and Analysis

4.1. Dataset. To evaluate the performance of our algorithm, we have chosen three widely used datasets. SED2 contains 100 nature images with two salient objects. Every image in the dataset was finely labeled manually for the purpose of saliency detection and image segmentation. ALL-IDB1 and ALL-IDB2 are the acute lymphoblastic leukemia image database [18].

4.2. Implementation Details. In this paper, input image with large size should be downsampled to 64*64 for fixation prediction and salient object detection, because SR algorithm is sensitive to image size, and 64 pixels of input image may be a good estimation of majority images. More importantly, reducing the size of image can save running time sharply. The number of superpixels of RBD could be set to 100 or 150. It is not sensitive to our method.

The number of positive and negative samples is set to 500 in sampling. And the number of hidden nodes of PELM may be linked to the dimension of pixels feature and could be set to 5~30 in this paper. In order to control the loop of positive feedback, F-measure is used to measure the similarity between BW_i and $BW\_i\text{-}1$. And $F = 0.95$ means both areas are similar enough in our experiment.

4.3. Evaluation Measures. We perform both quantitative and qualitative evaluations for our approach. For quantitative

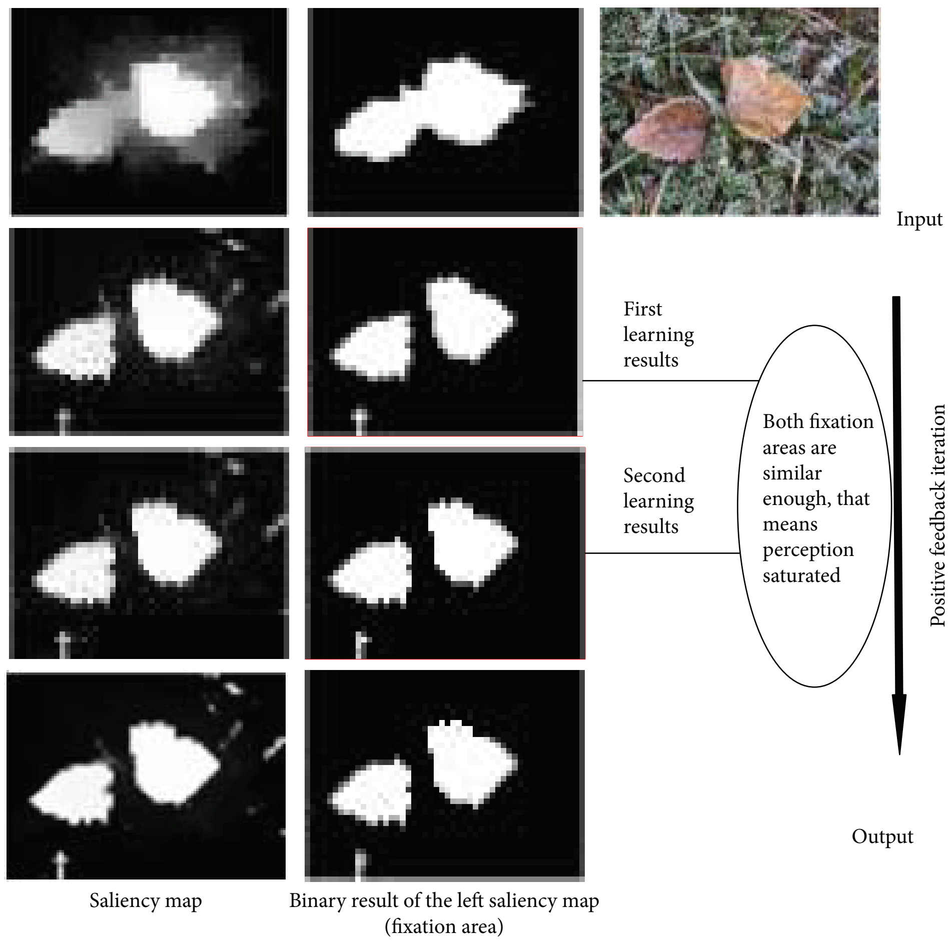


Figure 5: The positive feedback accelerates the process for the perception to become saturated.

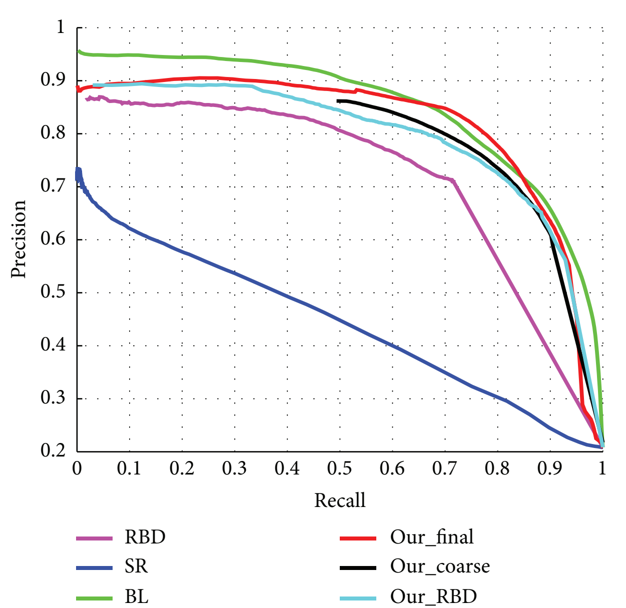


Figure 6: PR curves of ours and four compared algorithms in SED2.

evaluation, we use recall, precision, and *F*-measure. *F*-measure jointly considers recall and precision. For a saliency map *S*, we first covert it to a binary mask *M* by thresholding using a fixed threshold which changes from 0 to 255. On each threshold, a pair of P/R scores is computed to form PR-curve and to describe the performance of model at different situations. Recall and precision can be computed by the following function.

Table 1: Average *F*-measure of four compared approaches in SED2.

Method	SR	RBD	BL	Ours
F-measure	0.700	0.8250	0.8342	0.8561

$$P = \frac{|M \cap G|}{|M|}, \quad R = \frac{|M \cap G|}{|G|}, \tag{8}$$

where G denotes the ground truth, and *F*-measure can be defined as follows:

$$F = \frac{(1+\beta^2) * P * R}{\beta^2 * P + R}. \tag{9}$$

As suggested by the literature [8], β^2 is set to 0.3 to enhance the effect of precision. The more the F value is,

Figure 7: Comparison of partial experimental results in SED2.

Table 2: Average *F*-measure of different approaches in ALL-IDB1 and ALL-IDB2.

Method	ALL-IDB1 Watershed	ALL-IDB1 Reference [2]'s method	ALL-IDB1 Ours	ALL-IDB2 Watershed	ALL-IDB2 Reference [2]'s method	ALL-IDB2 Ours
F-measure	0.60	0.89	0.86	0.56	0.82	0.95

the better the performance. We take the average *F*-measure of each database as the final *F*-measure.

4.4. Experimental Results

4.4.1. Nature Image Saliency Detection and Segmentation. We firstly test our approach in SED2 database. Three models were compared which are state-of-the-art or closely related to our approach: BL [14], SR [9], and RBD [15]. PR curves of compared methods are shown in Figure 6. In PR curves, Our_final means saliency map output from the positive feedback loop; Our_coarse is the coarse saliency map output by first learning; Our_RBD is optimized saliency map after RBD. Other saliency maps are represented by algorithm names.

Figure 6 shows that the top-left of the BL's curve is higher than the others. It means that BL's saliency map is more detailed and smooth. However, Our_final is more

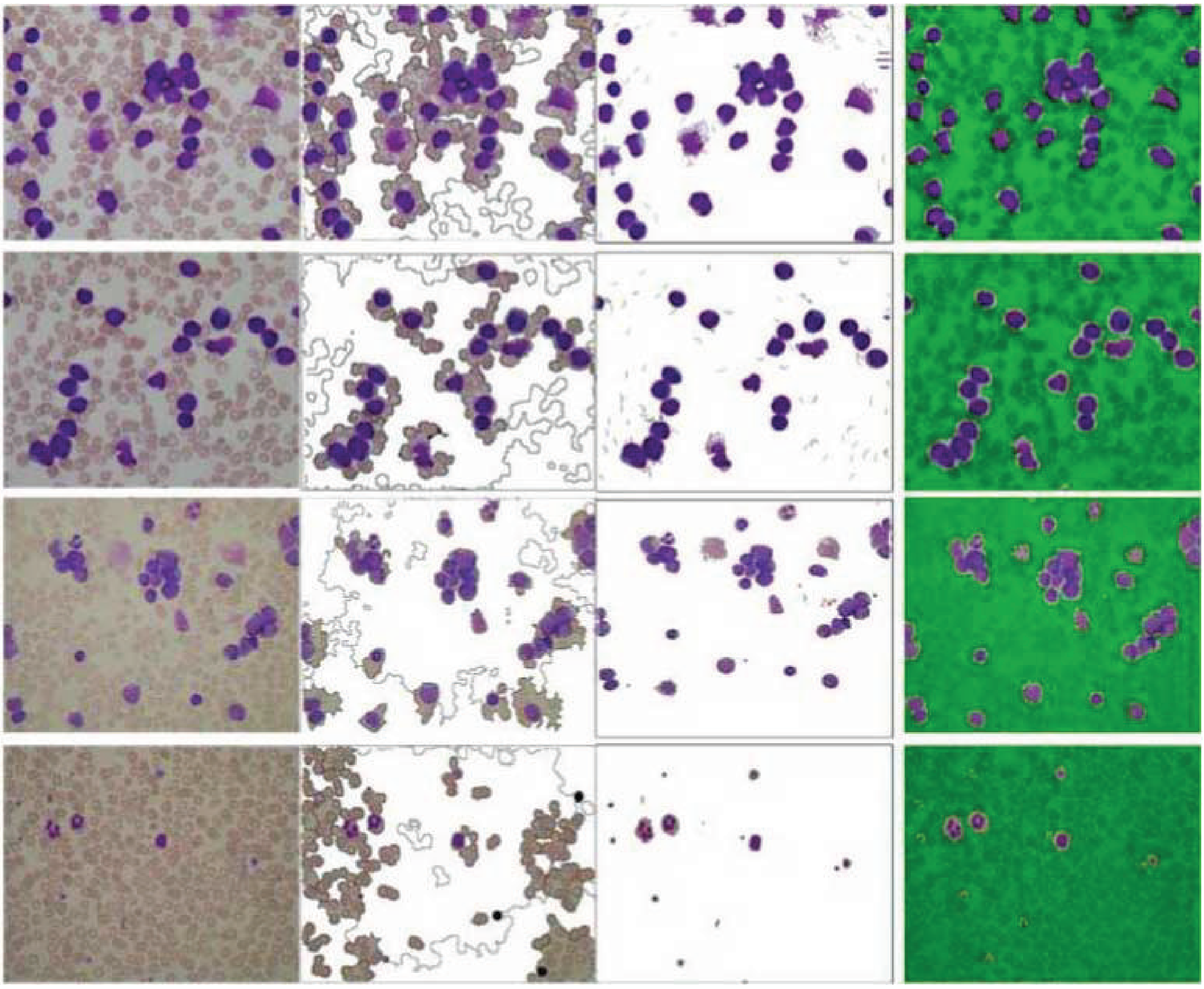


Figure 8: Comparison of partial experimental results in ALL-IDB1.

than BL in the middle of the PR curve that illustrates the ability of our method to grasp the whole object is better than BL. Besides, Our_coarse and Our_RBD are higher than the original RBD and SR. Although the curves of Our_coarse and Our_RBD are little lower than those of the BL, Our_final achieves good result after the positive feedback. Obviously, positive feedback has played a decisive role in improving performance.

Table 1 shows average F-measures for the 4 methods. Those results show that our method has the best performance, followed by BL and RBD, and SR is worst. It is also shown that the performance of SR and RBD can be improved effectively by adding learning-based positive feedback.

Figure 7 shows part of the images in the SED2 and their saliency maps obtained by the 4 methods. These results show that the BL saliency map is better in smoothness and detail, and our method is better in overall perception. It should be noted that SR, RBD, and our methods reduce the size of original image in saliency detection and their saliency maps are rougher than BL's one. From the view of qualitative evaluation, it is clear that the binary object mask detected by our method is closer to the ground truth.

4.4.2. Leukemia Image Segmentation. ALL-IDB1 contains 108 images with large field of vision, each image includes many WBCs. Some of them may overlap and touch together. ALL-IDB2 contains 260 images with small field of vision, and each of them only contains a nucleated cell. The difficulty lies in that conventional methods are hard to extract the entire leukocyte populations, due to the color of cytoplasm of WBCs often close to that of the background.

Two methods were compared with our approach: marker-controlled watershed and Reference [2]'s method. The former performs flooding operation according to the selected markers and the gradient. The latter firstly finds the deep stained nucleus of WBCs by thresholding and then does sampling around the fixation area and learning/classification by SVM/ELM. We sketched the outline of the nucleated cells in the image as ground truth. The average F-measures are shown in Table 2.

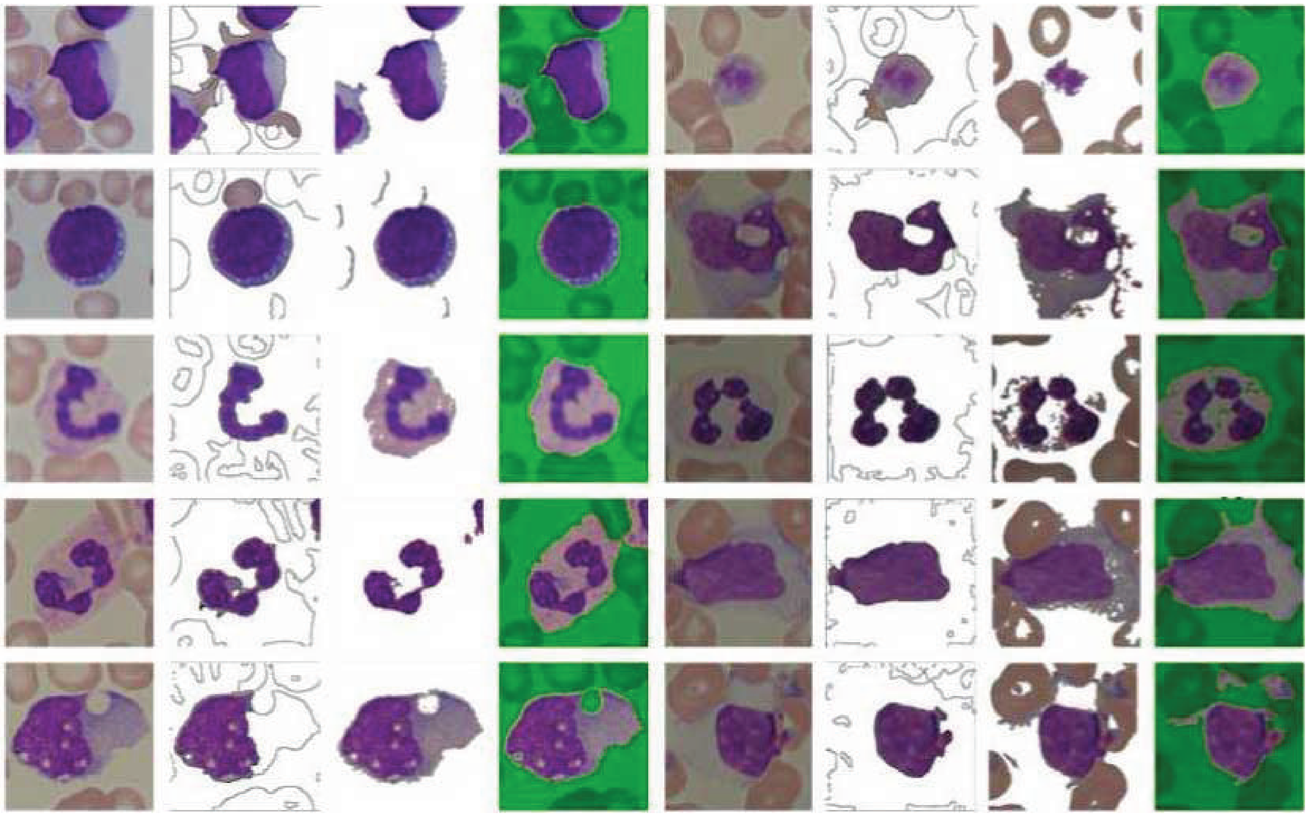


Figure 9: Comparison of partial poor results in ALL-IDB2.

Our method gets the highest score in ALL-IDB2, while slightly worse than Reference [2]'s method in ALL-IDB1. Watershed-based method is worst in both datasets.

Partial experimental results are listed in Figures 8 and 9. As can be seen from these examples, our method is successful in ALL-IDB2. In which only one stained cell exists, so the detail of the object may be well preserved in our method. In ALL-IDB1, a large number of stained cells gathered together may limit the performance of our method. While Pan's method without positive feedback procedure may be more appropriate to deal with this situation.

In ALL-IDB2, only one segmentation result is not ideal in our method. The segment result of this image losts most cytoplasm (shown in Figure 9, the right side of the last row); however, even the human eyes are prone to error in this image. These examples show that our approach is much close to human perception.

4.5. Discussion. The method of "pixel sampling-learning-classification" was proposed previously in [2]. It works well in good control condition. It needs be noted that the framework of the above method is very similar with step 2 in our method, in which shallow networks are parallelly arranged without any feedback. It is a noise sample-sensitive method if the training samples are not well prepared according to the prior knowledge. Our method could be regarded as an improved version developed from the technique of [2]. We presented an effective way to deal with the noise sample-sensitive problem by background-suppressing and learning-based positive feedbacks.

Our method also differs from Na's works especially in simulating human vision. Na's team tries to train a set of weak classifiers based on initial saliency maps and then obtains a strong model by integrating the weak classifiers. The final output relies on the strong model. They take boosting and parallel strategy to group weak classifiers, but without any feedback in their framework. In contrast, our method only focuses on the fixation region to accelerate the process for the object perception to become saturated, no matter how the classifier is weak or strong. In our method, the saliency map could be produced by accumulating the binary result in iteration and object could be output by thresholding the saliency map. By the way, multiscale analysis is not involved in our method. We just downsample image to a small size (64*64) that can sharply speed the algorithm, while it does not decrease the performance.

5. Conclusions

This paper proposes a novel saliency region detection method based on machine learning and positive feedback of perception. Motivated by human visual system, we construct a framework using EPELM to process visual information from coarse to fine, to form a saliency map and extract salient objects. Our algorithm is data-driven totally and needs no any prior knowledge compared with the existing algorithms.

Experiments on several standard image databases show that our method not only improves the performance of the conventional saliency detection algorithms but also segments nucleated cells successfully in different imaging conditions.

Conflicts of Interest

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

Acknowledgments

This research was supported by the Natural Science Foundation of Zhejiang Province of China (no. LY13F010004) and the National Natural Science Foundation of China (nos. 61672476 and 61462031).

References

[1] Z. Yu, H. S. Wong, and G. Wen, "A modified support vector machine and its application to image segmentation," *Image and Vision Computing*, vol. 29, no. 1, pp. 29–40, 2011.

[2] C. Pan, D. S. Park, Y. Yang, and H. M. Yoo, "Leukocyte image segmentation by visual attention and extreme learning machine," *Neural Computing and Applications*, vol. 21, no. 6, pp. 1217–1227, 2012.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, Boston, MA, USA, June 2015.

[4] X. Zheng, Y. Wang, G. Wang, and Z. Chen, "A novel algorithm based on visual saliency attention for localization and segmentation in rapidly-stained leukocyte images," *Micron*, vol. 56, pp. 17–28, 2014.

[5] R. Matrin, "Microsaccades: small steps on a long way," *Vision Research*, vol. 49, no. 20, pp. 2415–2441, 2009.

[6] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: a survey," *International Journal of Machine Learning and Cybernetics*, vol. 2, no. 2, pp. 107–122, 2011.

[7] J. W. Zhao, Z. H. Zhou, and F. L. Cao, "Human face recognition based on ensemble of polyharmonic extreme learning machine," *Neural Computing and Applications*, vol. 24, no. 6, pp. 1317–1326, 2014.

[8] A. Borji, M. M. Cheng, H. Z. Jiang, and J. Li, "Salient object detection: a benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[9] X. D. Hou and L. Q. Zhang, "Saliency detection:a spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, June 2007.

[10] S. Goferman, L. Z. Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[11] M. M. Cheng, N. J. Mitra, X. Huang, P. H. 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.

[12] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 717–729, 2016.

[13] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: unsupervised learning for object saliency and detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3238–3245, Portland, OR, USA, June 2013.

[14] T. Na, L. U. Huchuan, R. Xiang, and M.-H. Yang, "Salient object detection via bootstrap learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1884–1892, Boston, MA, USA, June 2015.

[15] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1911–1922, 2017.

[16] L. Zhang, J. Li, and H. Lu, "Saliency detection via extreme learning machine," *Neurocomputing*, vol. 218, no. 8, pp. 103–112, 2016.

[17] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2814–2821, Columbus, OH, USA, June 2014.

[18] R. D. Labati, V. Piuri, and F. Scotti, "ALL-IDB: the acute lymphoblastic leukemia image database for image processing," in *Proc. of the 2011 IEEE Int. Conf. on Image Processing (ICIP 2011)*, pp. 2045–2048, Brussels, Belgium, September 2011.

Hindawi
Journal of Healthcare Engineering
Volume 2018, Article ID 4098237, 13 pages
https://doi.org/10.1155/2018/4098237

Research Article

An Elderly Care System Based on Multiple Information Fusion

Zhiwei He, Dongwei Lu, Yuxiang Yang, and Mingyu Gao

College of Electronic Information, Hangzhou Dianzi University, Hangzhou, China

Correspondence should be addressed to Zhiwei He; zwhe@hdu.edu.cn

Received 20 February 2017; Revised 14 May 2017; Accepted 21 November 2017; Published 15 January 2018

Academic Editor: Yong Yang

With the development of social economy in the 21st century, and the rising of medical level, the aging of population have become a global trend. However lots of elderly people are in "empty nest" state. In order to solve the problem of high risk of daily life in this group, this paper proposed a method to integrate the information of video images, sound, infrared, pulse, and other information into the elderly care system. The whole system consists of four major components, that is, the main control board, the information acquisition boards, the server, and the client. The control board receives, processes and analyzes the data collected by the information acquisition boards, and uploads necessary information to the server, which are to be saved to the database. When something unexpected occurs to the elderly, the system will notify the relatives through the GPRS (general packet radio service) module. The system also provides an interface for the relatives to inquire the living status of the elderly through an app. The system can monitor the living status for the elderly with the characteristics of quick response, high accuracy, and low cost and can be widely applied to the elderly care at home.

1. Introduction

The aging of population is a global issue [1], especially in China. Most children who are busy with their work have little time to take care of their parents and have a great pressure on parent support. As most elderly become empty nesters, monitoring the living status of them is to solve not only family problems but also social problems. China's welfare science for elderly is still in the early stages of development [2, 3]. The existing products based on wearable sensors sometimes feel inconvenient and are easy to forget to be carried. The products based on audio sensor can judge the living condition of the elderly through sound signals, but they are vulnerable to environmental noise, which leads to low accuracy. The products based on vision sensors also have some problems such as limited visual acquisition and privacy leakage. Therefore, developing an elderly care and monitoring system which meets the privacy protection requirement has great significance in family, social, and practice value. The system should be able to effectively monitor the daily life and correctly assess the health status of the elderly. When something unexpected happens, the system will send an alarm signal to inform the family relatives or other related people.

Some related systems have been proposed in the literature [4–8]. For example, Kidd et al. [6] proposed the "Aware Home" system, which captured real-time images of the elderly through the camera, and the children can see the elderly current activity information through the Internet and can view the recorded information to better understand the status of the elderly. Recently, Khosla et al. [7] reported an interactive multimodal social robot system for improving quality of care of elderly in Australian nursing homes. In their system, they utilized multimodal interaction (voice, gestures, emotion, touch panel, and dance) in assistive social robot. Suryadevara and Mukhopadhyay [8] proposed a wireless sensor network-based home monitoring system for wellness determination of elderly. In their system, they used a number of sensors interconnected to detect usage of electrical devices, bed usage and chairs along with a panic button, and wireless sensor network consisting of different types of sensors like electrical and force, and contact sensors with Zigbee module sensing units are installed at elderly home.

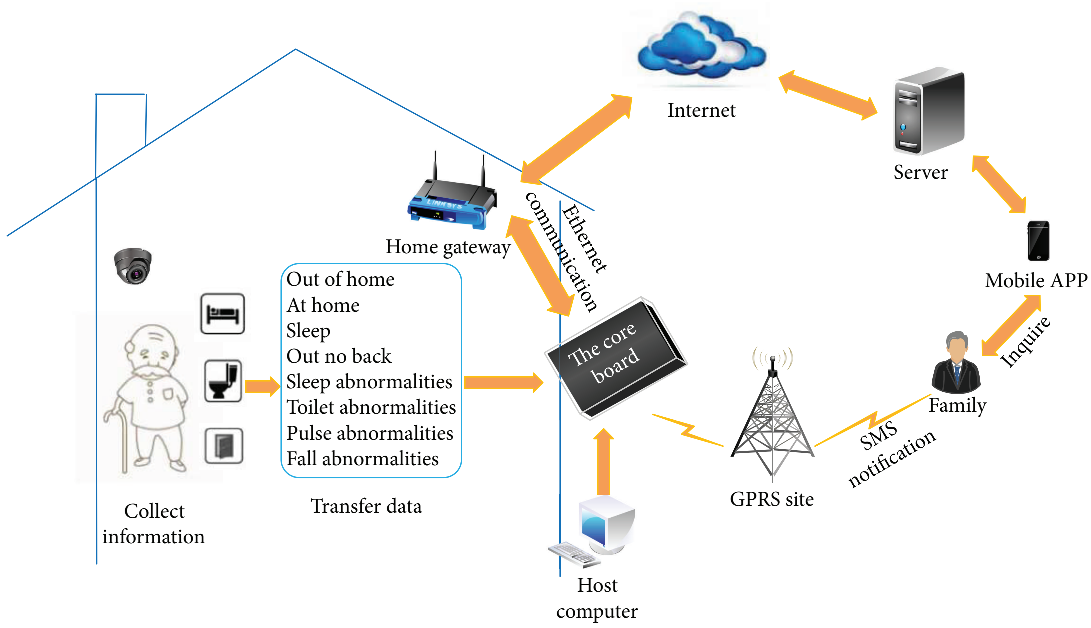


Figure 1: Schematic diagram of system.

In this paper, we propose an elderly care system based on multi-information fusion technology, using video processing technology as the core, combined with sound detection, infrared detection, and pulse detection. Specifically, the proposed system uses a DSP + ARM dual-core board, with OMAP L138 as the processor. A six-layer PCB is applied and designed in order to simplify the circuit and reduce the cost. Through the method of background modeling and updating, the foreground moving human being can be extracted properly with the help of an 8-connectivity analysis and shadow removal. Using some features of the minimum circumscribed rectangle, the falling-down of the elderly can be detected properly. In order to obtain the living information of the elderly without privacy disclosure, several information acquisition boards armed with infrared sensors, laser sensors, sound sensors, and pulse sensors are used at the gate, in the toilet, and in the bedroom. The multiple information fusion mechanism improves the efficiency and accuracy of the elderly care system. Hence, the developed system can accurately track the indoor position of the elderly, detect abnormal activities, and inform the relatives automatically when something unexpected happens. In general, this research undoubtedly provides important basis for the generalization and application of the elderly care system.

The rest of the paper is organized as follows. In Section 2, the whole system is described in detail. In Section 3, the video analysis-based falling-down detection algorithm is given. System setup and experimental results go to Section 4. Finally, in Section 5, the conclusions and discussions are given.

2. System Description

2.1. System Overview. The schematic diagram of the whole system is shown in Figure 1. The whole system consists of a main board and several information acquisition boards. The main control board is the core of the hardware system, while the information acquisition boards are the basis. The information acquisition boards are installed around the room at the right places. The voice, infrared, and pulse data are then collected directly by these information acquisition boards and some of the living status of the elderly, such as whether he/she is absent or is sleep abnormally, can then be obtained easily. The living status whether the elderly falls down is obtained through video analysis by the main board. When all these living conditions are obtained, they are then uploaded to the server through Ethernet. Relatives can view the real-time status and historical status of the elderly with his/her mobile phone through a special installed app on it. On the other hand, when something unexpected occurs and is detected by the system, a short message will be sent automatically to the relatives through a GPRS (general packet radio service) module installed on the system.

2.2. Design of the Hardware System. For the hardware design, three aspects should be considered. First, the hardware system works in an indoor environment, and the main influence is the temperature and weather change. Second, the cameras are fixedly installed, so the video analysis algorithm has a certain robustness to simple noise interference. Third, the system should work in real-time. According to these aspects, we build the hardware platform based on a DSP (digital signal processor) + ARM (acorn RISC machine) dual-core CPU (central processing unit) with OMAP (open multimedia application platform) L138 [9] as the processor which is developed by Texas Instruments, along with three infrared and sound detection modules (information acquisition control panel), one pulse detection module, and two analog cameras as the sensors. The hardware platform has powerful

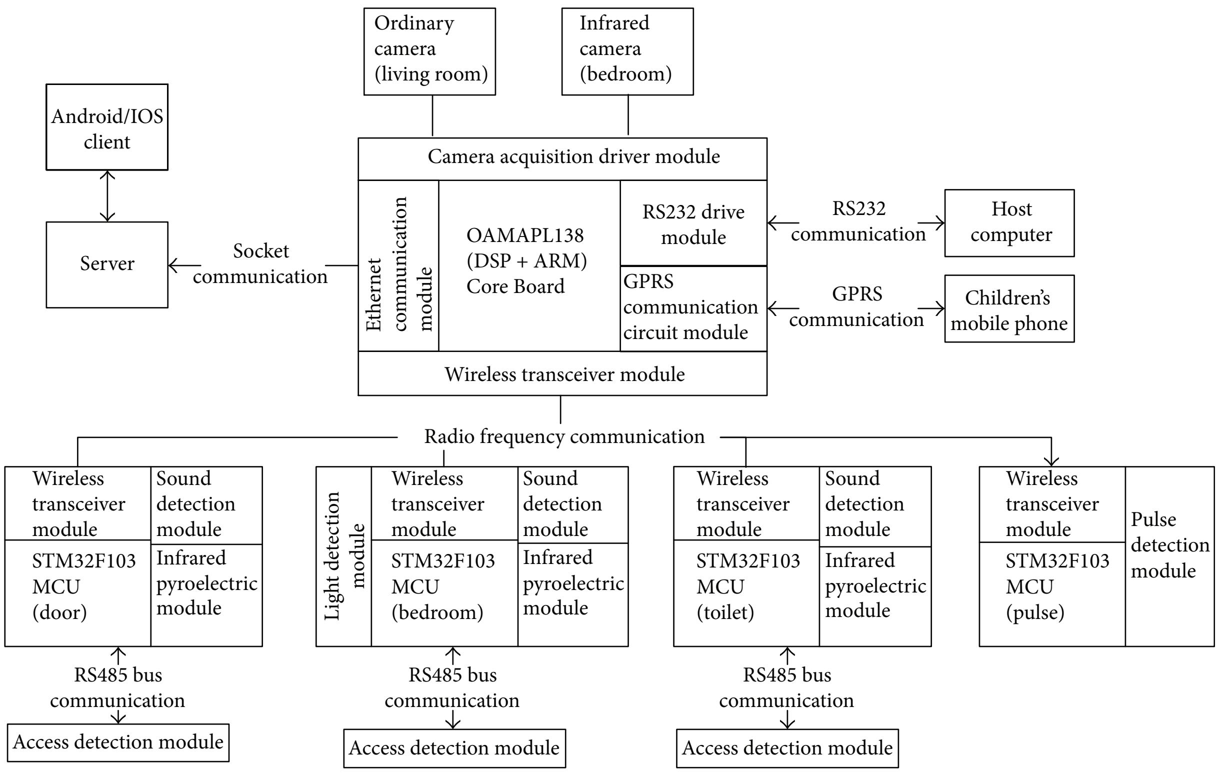


Figure 2: Schematic diagram of the hardware platform.

data processing capabilities to meet the system's real-time and efficiency requirements. Figure 2 shows the diagram of the system hardware platform.

(1) The core of the main board, OMAP L138, is a dual-core (DSP + ARM) CPU with an up to 456 MHz working frequency, a 512 MByte extended NAND Flash and a 128 MByte DDR2 (double data rate 2) memory, and a wealth of external interfaces such as the Ethernet, video, and LCD (liquid crystal display) interface, and so forth. The main control board utilizes the wireless transceiver module YB30_SI4432 which is developed by Silicon Labs (http://www.silabs.com) to communicate with the information acquisition board, which has the characteristics of long transmission distance, low cost, high integration, and high ability going through a wall. The living conditions and the captured images when the elderly falls down are sent through this module.

(2) There are four pieces of information acquisition board which uses the microcontroller STM32F103 [10], which is developed by ST Microelectronics, as the core. Three of them, armed with infrared, sound, and laser detection modules, are installed in the toilet, the bedroom, and the gate, respectively, to detect abnormal conditions including long-time staying in the toilet, not returning back home, or sleeping disorders. As mentioned before, once abnormality occurs, data will be transmitted to the main control board through the wireless transceiver module. The last one piece is used to detect the elder's pulse frequency. It can be carried with the elderly all the time or measured when needed.

(3) The main control board also judges abnormal status in the living room through an analog camera when the elderly falls down or detects abnormal sleeping status in the bedroom through another infrared camera.

(4) The dual-core main control board will send the collected information to the server through the Ethernet transmission. The server will save the data to a database and show the current status of the elderly.

2.3. Design of the Software System. In this paper, the software system consists of four parts: the main control board software, the information acquisition board software, the server software, and the client software. The functions of the main control board software include abnormal status judgment, notice warning, and fall detection. The functions of the information acquisition board software include infrared detection, laser detection, sound detection, and pulse detection. The main function of the server software is to do database operations, and the client software is the interface for relatives to view/review the status of the elderly. The diagram for the whole software is shown in Figure 3.

2.3.1. Design of the Main Control Board Software. The software of the main control board is divided into three layers from the bottom to the top: the peripheral driver function

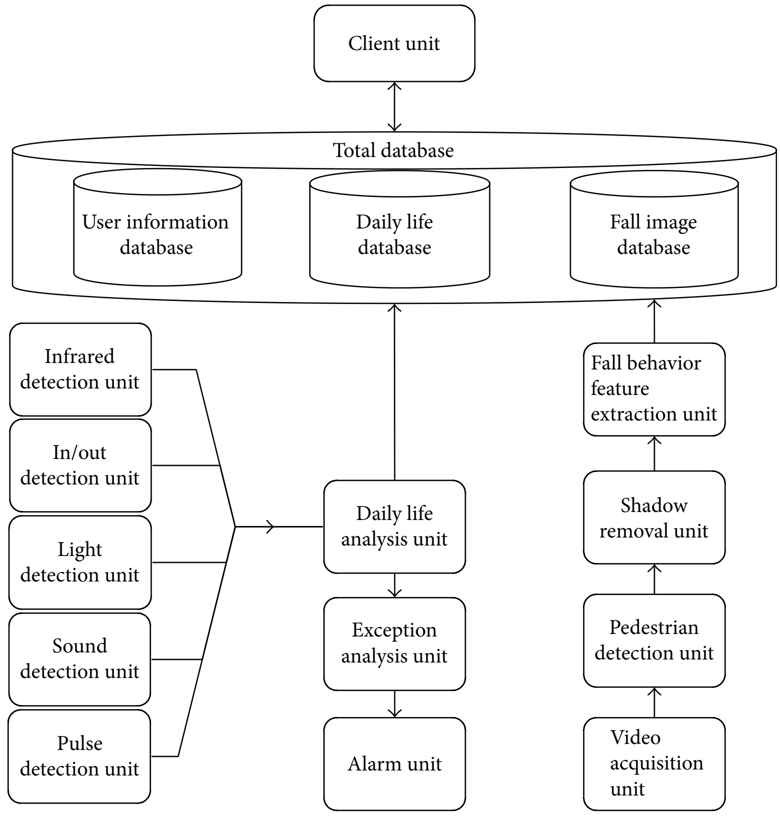


FIGURE 3: Schematic diagram of the system software.

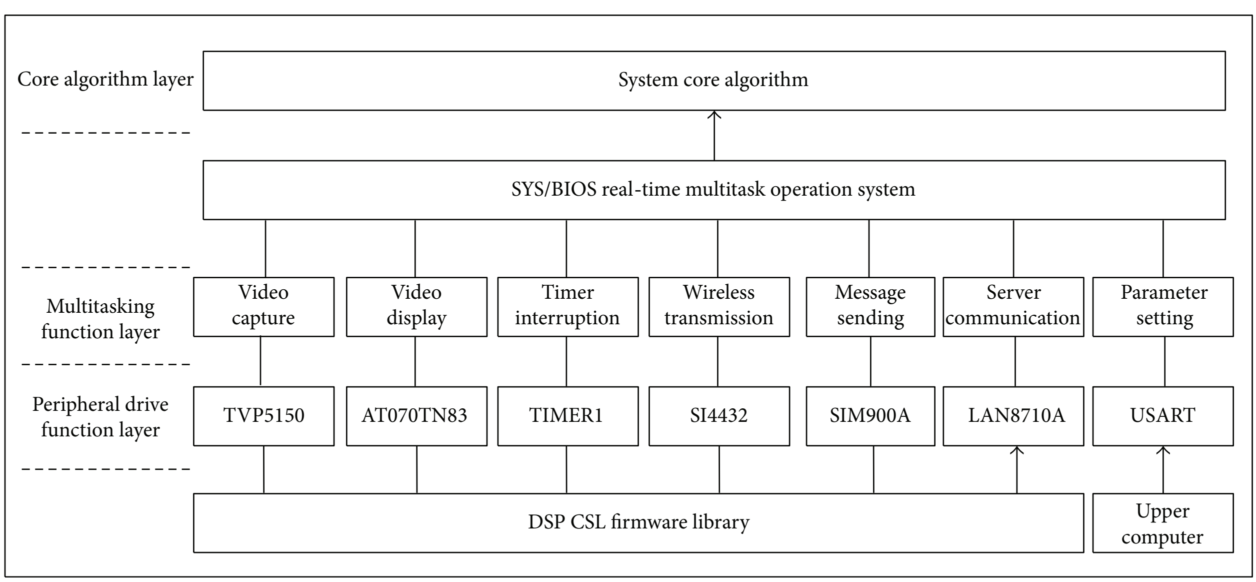


FIGURE 4: Software framework of the main control board.

layer, the multitask function layer, and the core algorithm layer. The driver function layer is used to initialize the peripheral of the main control board with CSL (chip support library) firmware library to drive them to work normally. The multitask function layer is designed to realize different tasks of the system, including the video capture, video display, timer interrupt, wireless transceiver, SMS sending, and server communication tasks. The core algorithm layer mainly accomplishes the fall detection of the elderly. The software framework of the main control board is shown in Figure 4.

According to the software framework and combined with the characteristics of the SYS/BIOS multitasking OS

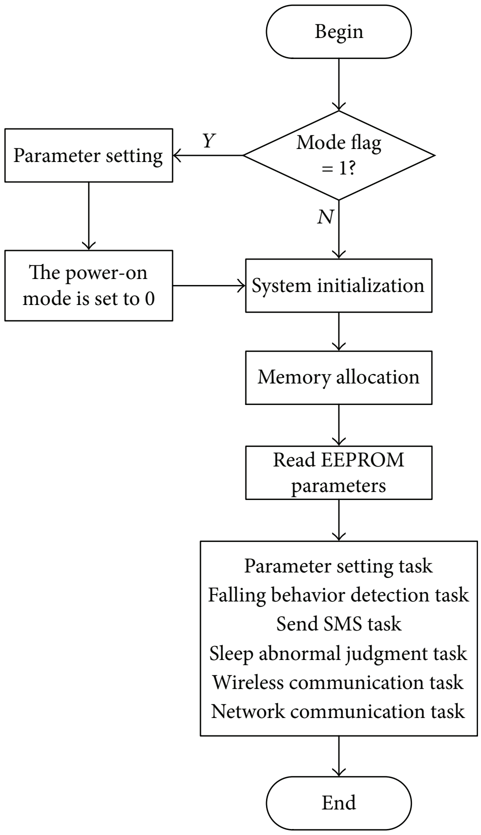


Figure 5: Software process of the main control board.

(operating system), the software workflow of the main control board is shown in Figure 5.

After the main board is powered on, the system will first do some initialization work such as system clock configuration, CSL library initialization, and peripheral and memory resource initialization. The details are as follows:

(1) System clock configuration: setting OMAP L138 system clock to 456 MHz

(2) Peripheral initialization: GPIO (general purpose input/output) port initialization, LAN8710A's driver initialization, EEPROM initialization, SI4432's driver initialization, TIMER1's driver initialization, TVP5150's driver initialization, AT070TN83's driver initialization, timer SIM900A's driver initialization, and RS232's driver initialization

(3) Memory resource management: mapping the appropriate data to DDR2

(4) Read and set the parameters in EEPROM

(5) Multitask operation: parameter setting tasks, falling behavior detection tasks (including target detection, target tracking, and falling behavior recognition), SMS task, sleep abnormality judgment task, wireless communication, and network communication task

The "Mode Flag" in Figure 5 is set through the hardware DIP switch. When the mode flag "ModeSet" is 1, the system is set at parameter setting mode, in which mode the manager can modify a series of parameters such as the room number, the mobile phone number and the threshold in the video analysis algorithm, or even the IP address through PC software. When the "ModeSet" is 0, the system is set at working mode.

2.3.2. Design of the Information Acquisition Board Software. As mentioned before, we use the information acquisition board to collect and judge abnormal living information at the gates, toilets, and bedrooms.

The software diagram of the information acquisition board is shown in Figure 6. It has three layers, that is, the peripheral driver function layer, the interface function layer, and the control algorithm layer. The peripheral driver function layer initializes and drives peripherals of the STM32. The timer TIMER2 is initialized as a general timer, which generates an interruption every 1 second to receive and determine the various signal and changes from sensors. The IN/OUT detection module communicates with the STM32 through the RS485 bus and received the IN/OUT status of the elderly through a serial port interrupt. The sound, light, and infrared sensors receive data through an external interruption and jointly judge the elderly living status. The SI4432 wireless module connects with the STM32 through the SPI bus interface for data communication. The board will send the elderly status information to the main control board.

The three living information of the elderly at the gate that should be detected include out (01), at home (02), and out without going home (03); the three living information of the elderly in the bedroom include getting up (04), sleeping (05), and sleep abnormality (06); the toilet abnormality (07) is the only a status that should be detected in the toilet. Each status is obtained through multiple sensors fusion.

Let us take the information collected at the bedroom as an example to explain the logic of the information fusion system:

(1) An entrance detection sensor detects that the elderly enters the bedroom.

(2) If the light sensor module detects that the elderly on the bed is moving, and the current time is the sleeping time of the elderly, the system judges that the elderly starts to sleep.

(3) When the elderly is sleeping, but the sound and infrared sensors cannot detect any effective data for more than 20 seconds, the system would suspect that sleep abnormality occurs.

(4) Within the sleeping mode, the entrance detection sensor module detects effective data, and if the current time is the wake-up time, the system would judge that the elderly gets up.

2.3.3. Design of the Client Software. The client software can be divided to two parts, the PC (personal computer) monitoring client and the mobile phone client.

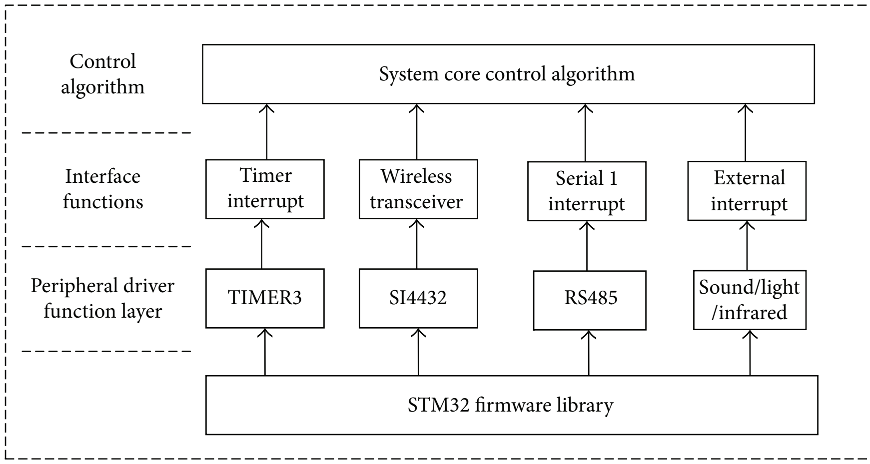


Figure 6: Software framework of the Information acquisition board.

(1) The PC monitoring client displays the data from the database server, including the living status of the elderly received from the main board and the abnormality along with the time of occurrence. The PC end software is developed in Visual Studio 2013 with the development language of C# and the database of SQLSever2012.

(2) In order to adapt to the most widely used two mobile phone operating systems, that is, the Android [11] and the iOS, two mobile phone clients are developed, respectively. The Android client software development platform is eclipse4.5, with the development language JAVA, and the iPhone client development platform is Mac OS X, with the language Objective-C. The mobile phone client includes a login interface, a status interface, and a message list interface. Users need to enter the correct username and password to login, which protect the privacy of the elderly to some extent. The server will record all the information into the database; when users log on to the mobile client to query the current status or images of the elderly, the server will package them in a standard Json format and send them to the mobile phone, and the mobile client will then parse the received package and print the message list for the users to view. If a falling-down event occurs, the relatives can view the falling-down image through the mobile phone client. The PC client and the mobile client communicate with each other according to the flowchart shown in Figure 7.

3. Video Analysis-Based Falling-Down Detection

The falling-down detection can be accomplished with many ways [12, 13]. In this paper, we utilize the video analysis based method for falling-down detection. Actually, video analysis-based object detection has been widely used in many areas [14].

Taking into account the privacy of the elderly, the cameras in the living room and the bedroom cannot capture real-time videos. Only when the elderly falls down, the abnormal images can be viewed through the mobile client. In this paper, the falling-down detection is divided into three steps: moving human-being detection, shadow removal, and falling-down feature extraction.

3.1. Moving Human Being Detection. We utilize the background subtraction method [15–17] for moving human being detection. There are mainly 3 steps for background subtraction-based moving object detection: background modeling, background updating, and background subtraction. In order to get a real-time moving human being detection, we propose an improved background modeling method; combined with the Surendra background updating algorithm [18], we can get a fast accurate detection.

3.1.1. Improved Background Modeling Algorithm. Background modeling refers to the extraction of the background from the video sequence, which is the key and basis step in the background subtraction algorithm. We propose an improved background modeling algorithm using the frame difference method. The core idea is to threshold the difference image to update the initial background (the initial gray value is 0) until the background is established. Considering the relatively slow motion of the elderly, the relative motion of the adjacent frames will be small or even static, which will lead to unsatisfactory results. So, in this paper, the two images doing difference operation are not adjacent frames which reflect the improvement. The core point of this algorithm is to compare the gray levels of the pixels at the same location in the two images at different times. When the threshold is less than a certain value, it is regarded as the background point. Specific steps are as follows:

Step 1. Take out five frames and every two frames have an F frame ($F = 15$ in this paper) separation in the original sequence. Initialize the gray value of the background image to be 0.

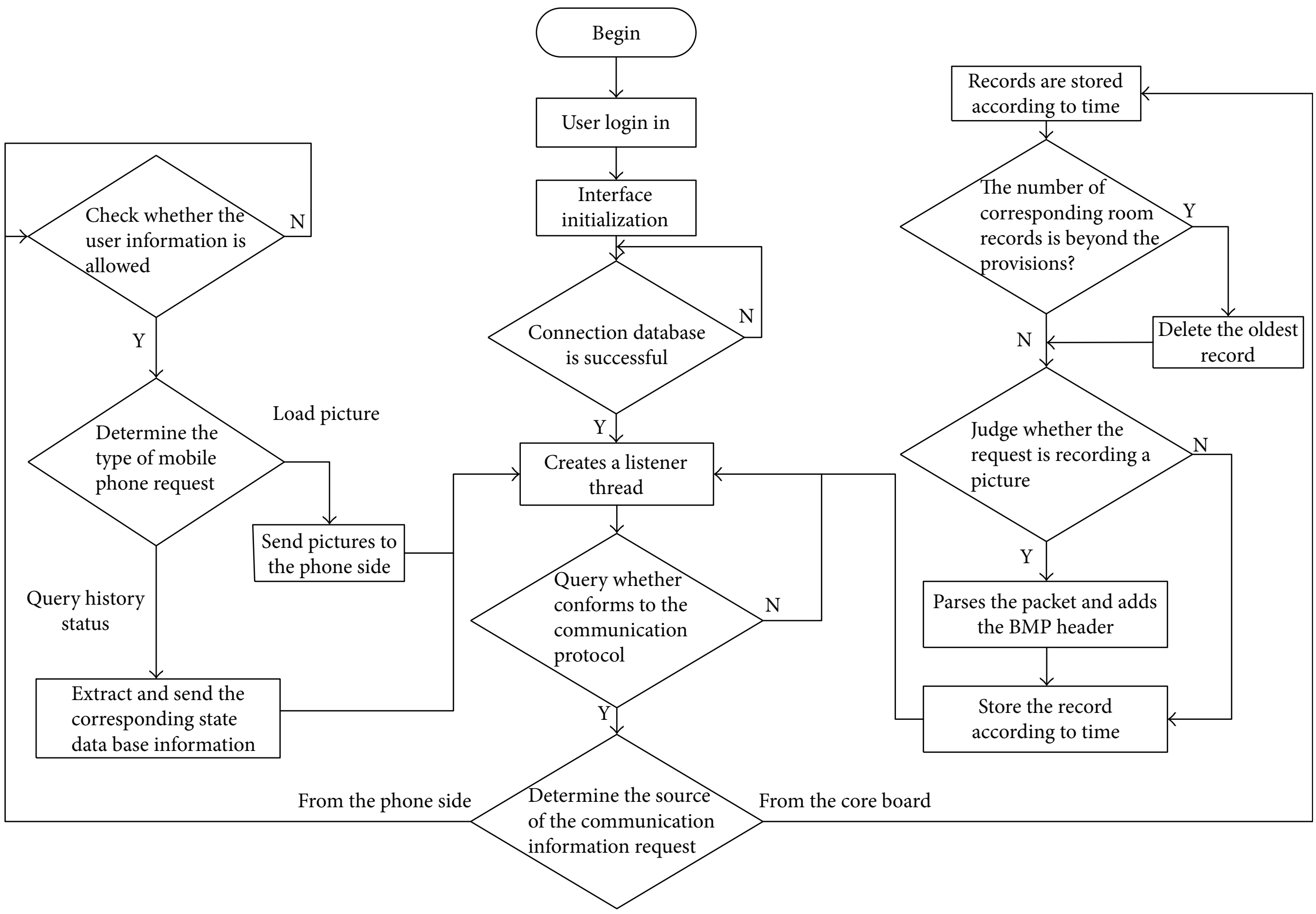


Figure 7: Client software flow chart.

Step 2. Obtain the difference image from the first frame and the second frame by image subtraction. If the gray value of a pixel in the difference image is less than a distinct threshold value T ($T = 20$ in this paper), and at the same time the gray value of this pixel in the background image is 0, set the gray value of this pixel in the background image to the same value as the second frames Hx.

$$\mathrm{BeiJingbuffer}(x,y) = \begin{cases} Hx, & |\mathrm{TempYbuffer2}(x,y) - \mathrm{TempYbuffer1}(x,y)| < \mathrm{T\&\&BeiJingbuffer}(x,y) = 0, \\ 0, & \text{else}. \end{cases} \tag{1}$$

Step 3. Repeat step 2 and take the second frame and the third frame images into the calculation, until all 5 images are completed, the background image is established.

Step 4. Update the established background image in real time: make a subtraction between the current image and the background image obtained by the third step, then get the difference image. If the value of a pixel in this difference image is greater than the threshold $T1$ (here is 25), then update the value to 255. We declare this point as a moving pixel and there is no need to update the background. Otherwise, the background image needs to update to 0, which is a binary image.

$$\mathrm{Rgb\_buffer} = \begin{cases} 255, & |\mathrm{Lumatopbuffer}(x,y) \\ & \quad - \mathrm{BeiJingbuffer}(x,y)| > T1, \\ 0, & \text{else}. \end{cases} \tag{2}$$

3.1.2. The Surendra Background Update Algorithm. The background template image is the initial background image of the previous $F * 5$ frame image, but the background image of the later frame is not static. Because of the influence of light and other objects, it is necessary to update the

	$(x+1,y)$	
$(x,y-1)$	(x,y)	$(x,y+1)$
	$(x-1,y)$	

(a) 4-neighborhood

$(x+1,y-1)$	$(x+1,y)$	$(x+1,y+1)$
$(x,y-1)$	(x,y)	$(x,y+1)$
$(x-1,y-1)$	$(x-1,y)$	$(x-1,y+1)$

(b) 8-neighborhood

Figure 8: Neighborhood positions.

background template in real time to adapt to the changes of indoor light and other environmental factors. Since this step requires a background image, only the still pixels need to be updated. In this paper, we use the following weight formula to update the background, that is, the Surendra background update algorithm.

$$\mathrm{BeiJingbuffer}(x,y) = \alpha \times \mathrm{Lumatobuffer}(x,y) + \beta \times \mathrm{BeiJingbuffer}(x,y). \quad (3)$$

$\mathrm{BeiJingbuffer}(x,y)$ is one pixel of the background image BeiJingbuffer, and $\mathrm{Lumatopbuffer}(x,y)$ is one pixel of the current image Lumatopbuffer. A and B are the weight coefficients, which meet $\alpha+\beta=1$, and inequality $\alpha\le 0.5$ and $\beta\ge 0.5$. α and β are used to adjust the background update speed. When α gets larger, the background image BeiJingbuffer will adapt to the scene more quickly but will cause larger foreground noise. When α gets lower, BeiJingbuffer will adapt more slowly and will cause lower noise. Therefore, to a certain extent, α weakens the effect of foreground pixels, and β enhances the role of the static background pixels. In this paper, we take experience values $\alpha=0.2$ and $\beta=0.8$.

3.1.3. Neighborhood Connectivity Analysis. Connected neighborhoods are neighboring pixels that are connected to each other by some similar rule, which is often a separate area of pixels in the image. In binary maps, there are only two-pixel values of 0 and 1, so this rule is usually specified by comparing the pixel values of neighboring pixels. Commonly used neighborhood connectivity domains are 4-neighborhood and 8-neighborhood. 4-neighborhood contains the positions at the top, bottom, left, and right positions of the target pixel, which is denoted as N_4 (q), as shown in Figure 8(a). 8-neighborhood contains the positions at the top, bottom, left, and right and the four diagonals of the target pixel, which is denoted as N_8 (q), as shown in Figure 8(b).

In this paper, we use the 8-neighborhood for connectivity analysis, and the concrete steps are as follows:

Step 1. Scan the binary foreground image line by line and find the first foreground pixel as the target pixel, then make a mark N_1.

Step 2. Label the same mark N_1 for all the pixels in the 8-neighborhood area of the foreground image and do the same operation for these 8 pixels until no more foreground pixels can be found.

Step 3. Scan to the next target pixel, if this pixel has been marked, then skip to the next pixel in the same 8-neighborhood area without any operation; otherwise, make a marker N_2 which is unused. Repeat step 2.

Step 4. Repeat step 3 until all the pixels in the image are scanned.

The operation of the above steps can get M label values $N_1, N_2, \ldots, N_M$ in the image; then, there are M-independent 8-connected neighborhoods. Count the size of these areas and arrange them in descending order to eliminate the area that is smaller than some distinct threshold. The background pixel values of these eliminated areas are set to 0.

Several examples are shown in Figure 8 for moving human being detection. In Figure 8, there are totally four sequences. Sequences (a) and (b) have the same background under normal light conditions, but people wear clothes with different colors. Sequence (c) was captured at the laboratory corridor open space, which has white walls and light-colored tiles and the background is relatively simple, but the light is sufficient. Sequence (d) is captured near the laboratory console, which has a very complex background. The first column of Figure 9 shows the original image of the sequences, the second column shows the foreground images extracted by the Surendra background update algorithm, and the third column is the results of the foreground after the 8-neighborhood analysis. For sequences (a), (b), and (c), the moving people are detected perfectly, but shadows are also detected due to the light reflection. For sequence (d), much more noise exist in the foreground after background subtraction, the reason is that the background and the light condition are much more complex. From (a) and (b), we can notice that the dress color has little effect on the moving object detection. From the second column, we can still see that a small amount of interfering pixels exist in the segmented foreground; this is due to the environmental noise and shadowing of human motion, but they can be eliminated perfectly after an 8-neighborhood connectivity analysis, as can be seen from the third column.

3.2. Shadow Removal Based on HSV Color Space. After the moving object detection, the shadow caused by human occlusion and light change still exists in the binary foreground image, and it cannot be removed by the connected domain analysis. It is necessary to separate the moving human foreground pixels from the shadow pixels so as to avoid interference with the subsequent extraction of the falling-down

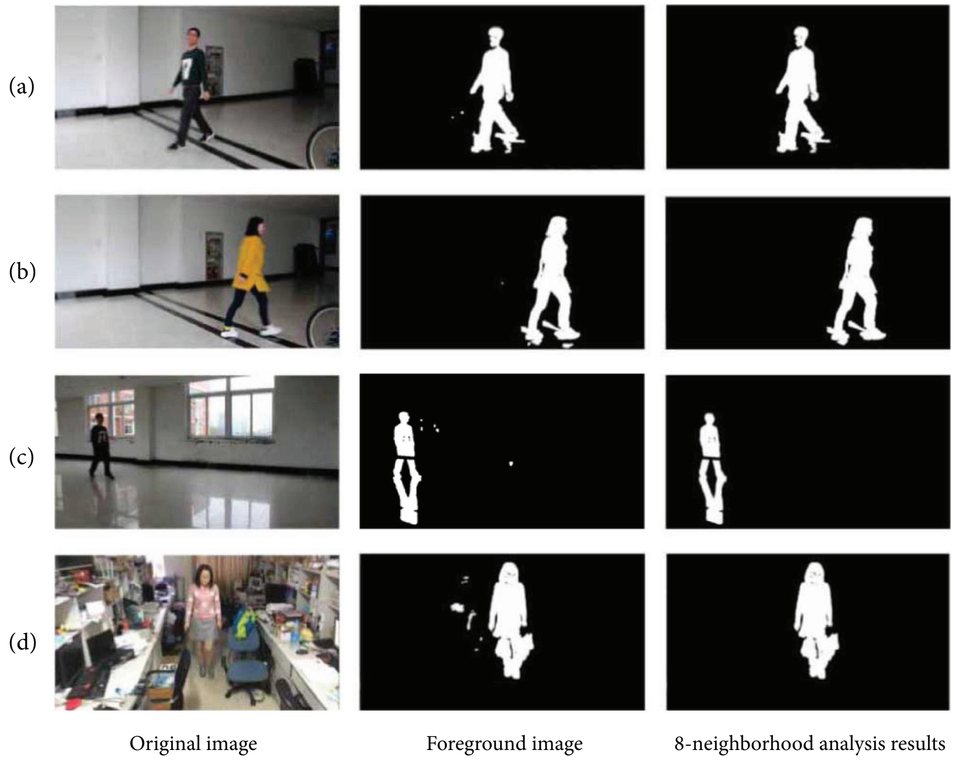


Figure 9: Foreground objects detection.

feature. In this paper, we do the shadow removal [19–22] in the HSV color space. The HSV color space is the most commonly used color model in machine vision research. The use of the HSV color space for shadow removal is intuitive: firstly, the shadow is often caused by light change or object occlusion, so the brightness component V of a shadow pixel is almost lower than both of the background pixels and the foreground pixels at the same location; secondly, the color tone component H of a shadow pixel is almost constant; and thirdly, the saturation component S of the shadow is almost low.

For indoor applications, the cameras are fixedly installed, so the background is static. In this case, the reflection coefficient of the detected target shadow points $\rho_t(x, y)$ is equal to that of the background points $\rho_B(x, y)$, as depicted in

$$\rho_B(x, y) = \rho_t(x, y). \tag{4}$$

On the other hand, the pixel brightness $S(x, y)$ is a product of the light intensity $E(x, y)$ and the reflection coefficient $\rho(x, y)$, so we have

$$S(x, y) = E(x, y)\rho(x, y). \tag{5}$$

According to (4) and (5), we can obtain the brightness ratio $\mathrm{RE}_t(x, y)$ between the shadow pixel and the background pixel as follows:

$$\mathrm{RE}_t(x, y) = \frac{S_t(x, y)}{S_B(x, y)} = \frac{E_t(x, y)}{E_B(x, y)}. \tag{6}$$

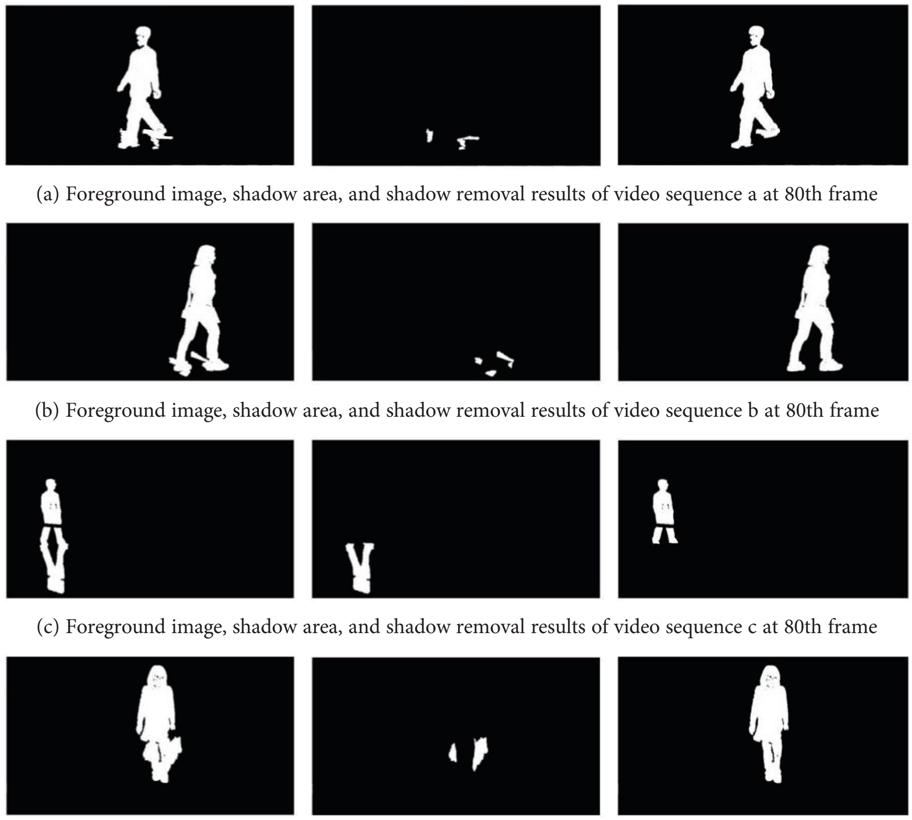
According to the principle of shadow optics, the light intensity $E(x, y)$ can be obtained with

$$E(x, y) = \begin{cases} L_A + L_P \times \cos\angle(\mathbf{n}(x, y), J), & \text{lighting}, \\ L_A, & \text{shadow}, \end{cases} \tag{7}$$

where L_A is the light source intensity, J is the light source direction, and $\mathbf{n}$ is the surface normal vector. It can be concluded from the above analysis that $\mathrm{RE}_t(x, y) \le 1$, which is in line with people's visual understanding of the shadow.

In the paper, we utilize the HSV shadow elimination algorithm (8) to detect the shadow points.

$$\mathrm{Shadow}(x, y) = \begin{cases} 1, & \lambda \le \dfrac{F_V(x, y)}{B_V(x, y)} \le \delta \cap (F_S(x, y) - B_S(x, y)) \le \alpha_S \cap |F_H(x, y) - B_H(x, y)| \le \alpha_H, \\ 0, & \text{other}, \end{cases} \tag{8}$$



(a) Foreground image, shadow area, and shadow removal results of video sequence a at 80th frame

(b) Foreground image, shadow area, and shadow removal results of video sequence b at 80th frame

(c) Foreground image, shadow area, and shadow removal results of video sequence c at 80th frame

(d) Foreground image, shadow area, and shadow removal results of video sequence d at 80th frame

FIGURE 10: Experiment results of the shadow removal.

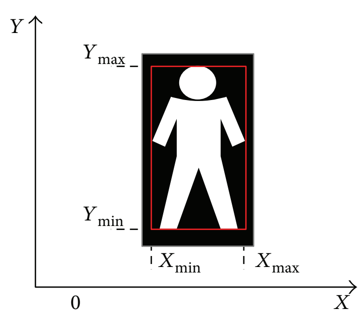


FIGURE 11: The sketch of the minimum circumscribed rectangle.

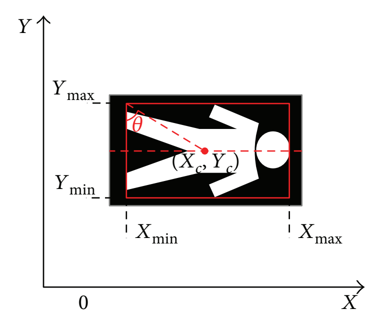
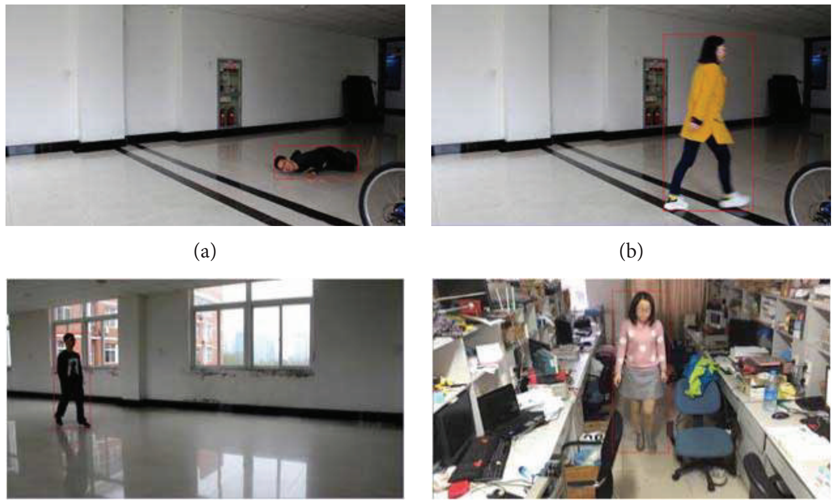


FIGURE 12: The minimum circumscribed rectangle in a falling-down state.

where $F_H(x, y)$, $F_S(x, y)$, and $F_V(x, y)$ are the H, S, and V components of the current image $F(x, y)$; $B_H(x, y)$, $B_S(x, y)$, and $B_V(x, y)$ are the H, S, and V components of the background image $B(x, y)$; and Shadow(x, y) is a binary image that shows whether a pixel is a shadow point. The greater the ambient brightness is, the smaller λ is; δ is a parameter set to avoid too many points being mistaken for shadow points to enhance robustness. λ and δ satisfy $0 < \lambda < \delta < 1$. The shadows are more saturated and color change is not obvious, so $1 > \alpha_S > 0$. In order to test the results more satisfactory, a value α_H was added to the limit, which can be adjusted according to specific application scenarios.

The effect of shadow removal is shown in Figure 10. The video sequences are the same to those in Figure 8. In Figure 9, the first column is the result after the 8-neighborhood connectivity analysis. The second column is the detected shadow, and the third column is the final foreground images after the shadow removal. We can see that the shadows are removed perfectly.

3.3. Falling-Down Feature Extraction. A falling-down means that the elderly falls down suddenly and does not stand up by himself in a period of time. In this paper, we use



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

Figure 13: Example obtained minimum circumscribed rectangles.

the minimum circumscribed rectangle for falling-down feature extraction.

The minimum circumscribed rectangle is the rectangle including the smallest area of all the points in a certain area. For falling-down detection, this area is just the segmented foreground human body. The sketch of the circumscribed rectangle is shown in red in Figure 11.

In Figure 11, four parameters $X_{\max}$, $X_{\min}$, $Y_{\max}$, and $Y_{\min}$ are needed when plotting the rectangle, which are the minimum and maximum coordinates along the x-axis and the y-axis. As can be seen from Figure 11, $X_{\min}$ is the x position of the leftmost point in an 8-connected region, while $X_{\max}$ is the rightmost point in an 8-connected region. Similar meanings are with $Y_{\max}$ and $Y_{\min}$.

As shown in Figure 12, when the elderly falls down, the aspect ratio of the minimum circumscribed rectangle changes rapidly, so this aspect ratio can be used as a feature for falling-down detection. But if the elderly is too close or too far away from the camera, using the aspect ratio as the criterion will lead to a failure of detection. On the other hand, when the elderly falls, the biggest change is the center of gravity in the x-axis direction, while the y-axis direction of the center of gravity does not change much. So, in this paper, we combine the aspect ratio K, the absolute slope of the center of mass $|S|$, and the center of gravity in the x-axis direction X_{mid} to determine whether a falling-down occurs, which are calculated as follows:

(1) The aspect ratio K:

$$K = \frac{Y_{\max} - Y_{\min}}{X_{\max} - X_{\min}}. \tag{9}$$

(2) The slope of the centroid S:

$$S = \frac{X_c - X_{\min}}{Y_c - Y_{\min}}, \tag{10}$$

where X_c and Y_c are defined in

$$X_c = \frac{\sum_{x=X_{\min}}^{X_{\max}} x^* n_x}{n}, \quad Y_c = \frac{\sum_{y=Y_{\min}}^{Y_{\max}} y^* n_y}{n}. \tag{11}$$

In (11), n_x is the number of foreground pixels in the xth column of the object, n_y is the number of foreground pixels in the yth row, and n is the total number of foreground pixels.

(3) The center of gravity in the x-axis direction X_{mid}:

$$X_{\mathrm{mid}} = \frac{\sum_{(x,y)\in H} x^* H(x,y)}{\sum_{(x,y)\in H} H(x,y)}. \tag{12}$$

In (12), $H(x, y)$ is the gray value of the moving object in the rectangle frame at pixel (x, y).

When the elderly falls, $K < 1$, $|S| > 1$, and X_{mid} has a big change.

Figure 13 shows several examples of the minimum circumscribed rectangles obtained according to the method proposed in this paper.

4. System Setup and Experimental Results

According to the functions of the elderly care system, we designed and implemented the whole system. As mentioned before, the system is divided into two components, the main control board and the information acquisition board. The designed embedded platform is shown in Figure 14. In Figure 14, the right-hand side is the main control board, which is used for video analysis and multi-information fusion. The left-hand side is the information acquisition boards, which are placed at the door, the toilet, and the bedroom to collect information of the elderly living information.

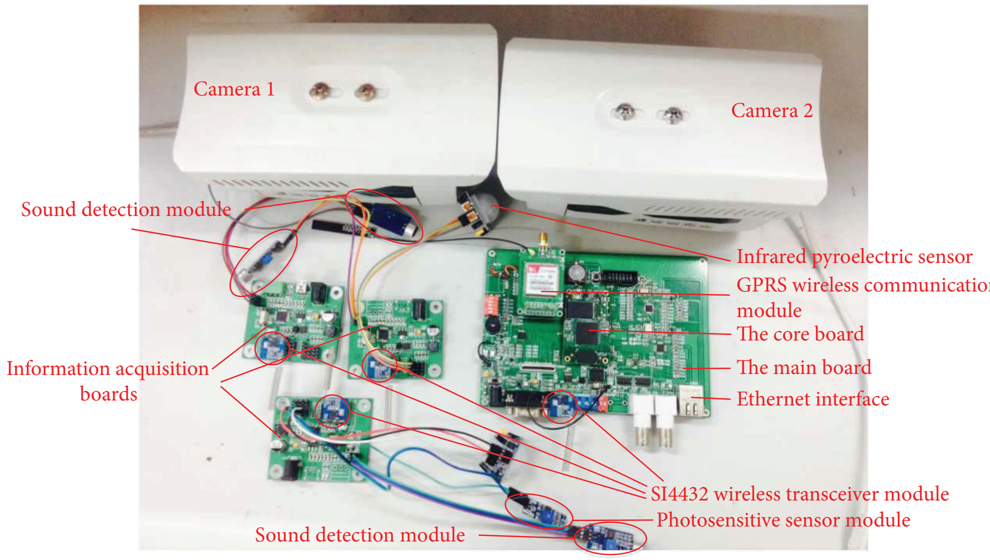


Figure 14: The designed system and its setup.

Table 1: System function test results.

Function	Total test numbers	Correct result	Accuracy rate
Out	100	100	100%
Out without back	150	150	100%
At home	100	100	100%
Toilet abnormality	150	147	98%
Sleeping	100	99	99%
Sleeping abnormality	150	146	97.33%
Getting up	100	100	100%
Falling-down	150	142	94.6%

The hardware platform, armed with the previously described algorithms, is thoroughly tested. The experimental results are shown in Table 1.

As can be seen from Table 1, eight functions of the system were tested. In Table 1, the status "out," "out without back," "at home," and "toilet abnormality" are obtained easily with the infrared pyroelectric sensor modules, through the detected moving direction of the elderly. Here, "Toilet abnormality" means that the elderly stays too long in the toilet. The "sleeping" and "sleeping abnormality" are obtained with the sound detection module mounted on the bed. A "sleeping abnormality" status means that the elderly may not breathe for a certain period of time. The "getting up" status and "falling-down" status are obtained through video analysis, as aforementioned. Experimental results show that the system meets the design requirement.

5. Conclusions and Discussions

In this paper, we propose an elderly care system based on multiple information fusion. Experiments demonstrate that the system can effectively notify the relatives through the GPRS when something unexpected occurs to the elderly. Moreover, the system can provide an interface for the relatives to inquire the living status of the elderly through an app installed on their mobile phone. In general, the developed system has the characteristics of quick response, high accuracy, and low cost, which can meet the requirements of real-time monitoring of the living status for the elderly and can be widely applied to the elderly care at home. In the future work, the proposed system can be further improved by integrating with other wearable sensors, for example, the ECG sensor [23], which can give much more information about the living status of the elderly.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the National Nature Science Foundation of China with Grant no. 61401129 and the Natural Science Foundation of Zhejiang Province (LY17F010020).

References

[1] A. Wong, *Population Aging and the Transmission of Monetary Policy to Consumption*, 2015, https://economics.stanford.edu/sites/default/files/arlene_wong_jmp_latest-2g9f9ga.pdf.

[2] B. Ma, *The Monitor System of Elderly People Living Alone Based on the Comprehensive Computer Vision*, Zhejiang University of Technology, 2014.

[3] Y. Bai, J. Li, and J. He, "The design of the fall detection system based on embedded video monitoring," *Television Technology*, vol. 38, no. 15, 2014.

[4] L. Liu, E. Stroulia, I. Nikolaidis, A. Miguel-Cruz, and A. Rios Rincon, "Smart homes and home health monitoring

technologies for older adults: a systematic review," *International Journal of Medical Informatics*, vol. 91, pp. 44–59, 2016.

[5] A. Jacobsson, M. Boldt, and B. Carlsson, "A risk analysis of a smart home automation system," *Future Generation Computer Systems*, vol. 56, pp. 719–733, 2015.

[6] C. D. Kidd, R. Orr, G. D. Abowd et al., *The Aware Home: a Living Laboratory for Ubiquitous Computing Research International Workshop on Cooperative Buildings*, Springer, Berlin Heidelberg, 1999.

[7] R. Khosla, M. T. Chu, R. Kachouie, K. Yamada, F. Yoshihiro, and T. Yamaguchi, "Interactive multimodal social robot for improving quality of care of elderly in Australian nursing homes," in *Proceedings of the 20th ACM international conference on Multimedia - MM '12*, pp. 1173–1176, New York, NY, USA, October-November 2012.

[8] N. K. Suryadevara and S. C. Mukhopadhyay, "Wireless sensor network based home monitoring system for wellness determination of elderly," *IEEE Sensors Journal*, vol. 12, no. 6, pp. 1965–1972, 2012.

[9] B. Sathiyabama and S. Malarkkan, "Low power adders for MAC unit using dual supply voltage in DSP processor," *International Proceedings of Computer Science & Information Tech*, vol. 32, p. 67, 2012.

[10] Z. Luo, Z. Liu, J. Zhang, and C. Song, "The design of musical instrument tuning system based on stm32f103 microcomputer," in *Proceedings of 2012 International Conference on Measurement, Information and Control*, pp. 79–82, Harbin, China, May 2012.

[11] Y. N. Zhang, H. Y. Ning, J. Bai, B.-C. Chen, P.-C. Zhou, and X.-L. Zhao, "Elderly safety early-warning system based on android mobile phones," in *2014 10th International Conference on Natural Computation (ICNC)*, pp. 1126–1130, Xiamen, China, August 2014.

[12] C. F. Lai, Y. M. Huang, J. H. Park, and H. C. Chao, "Adaptive body posture analysis for elderly-falling detection with multisensors," *IEEE Intelligent Systems*, vol. 25, no. 2, pp. 20–30, 2010.

[13] S. H. Kim and D. W. Kim, "A study on real-time fall detection systems using acceleration sensor and tilt sensor," *Sensor Letters*, vol. 10, no. 5, pp. 1302–1307, 2012.

[14] N. Buch, S. A. Velastin, and J. A. Orwell, "Review of computer vision techniques for the analysis of urban traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 3, pp. 920–939, 2011.

[15] K. Bahadir and K. Serdar, "Moving object detection and tracking by using annealed background subtraction method in videos: performance optimization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 33–43, 2012.

[16] M. Cheng and J. Gao, *An Improved Background Modeling Method for Target Detection*, Springer, Berlin Heidelberg, 2012.

[17] P. Gorur and B. Amrutur, "Speeded up Gaussian mixture model algorithm for background subtraction," in *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 386–391, Klagenfurt, Austria, August-September 2011.

[18] S. Gupte, O. Masoud, R. F. K. Martin, and N. P. Papanikolopoulos, "Detection and classification of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 1, pp. 37–47, 2002.

[19] A. Yoneyama, C. H. Yeh, and C. C. J. Kuo, "Moving cast shadow elimination for robust vehicle extraction based on 2D joint vehicle/shadow models," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, 2003*, p. 229, Miami, FL, USA, July 2003.

[20] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 59–68, 2006.

[21] A. Leone, C. Distante, N. Ancona, E. Stella, and P. Siciliano, "Texture analysis for shadow removing in video surveillance systems," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, pp. 6325–6330, The Hague, Netherlands, October 2004.

[22] R. W. Rahmat, Z. H. Al-Tairi, M. I. Saripan, and P. S. Sulaiman, "Removing shadow for hand segmentation based on background subtraction," in *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, pp. 481–485, Kuala Lumpur, Malaysia, November 2012.

[23] J. Yoo, L. Yan, S. Lee et al., "A wearable ECG acquisition system with compact planar-fashionable circuit board-based shirt," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 897–902, 2009.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 6506049, 11 pages
https://doi.org/10.1155/2017/6506049

Research Article

An Improved Random Walker with Bayes Model for Volumetric Medical Image Segmentation

Chunhua Dong,[1] Xiangyan Zeng,[1] Lanfen Lin,[2] Hongjie Hu,[3] Xianhua Han,[4] Masoud Naghedolfeizi,[1] Dawit Aberra,[1] and Yen-Wei Chen[2,4]

[1]*Department of Mathematics and Computer Science, Fort Valley State University, Fort Valley, GA, USA*
[2]*College of Computer Science and Technology, Zhejiang University, Hangzhou, China*
[3]*Radiology Department, Sir Run Run Shaw Hospital, Medical School of Zhejiang University, Hangzhou, China*
[4]*Graduate School of Information Science and Engineering, Ritsumeikan University, Kyoto, Japan*

Correspondence should be addressed to Yen-Wei Chen; chen@is.ritsumei.ac.jp

Received 24 February 2017; Accepted 23 April 2017; Published 23 October 2017

Academic Editor: Pan Lin

Random walk (RW) method has been widely used to segment the organ in the volumetric medical image. However, it leads to a very large-scale graph due to a number of nodes equal to a voxel number and inaccurate segmentation because of the unavailability of appropriate initial seed point setting. In addition, the classical RW algorithm was designed for a user to mark a few pixels with an arbitrary number of labels, regardless of the intensity and shape information of the organ. Hence, we propose a prior knowledge-based Bayes random walk framework to segment the volumetric medical image in a slice-by-slice manner. Our strategy is to employ the previous segmented slice to obtain the shape and intensity knowledge of the target organ for the adjacent slice. According to the prior knowledge, the object/background seed points can be dynamically updated for the adjacent slice by combining the narrow band threshold (NBT) method and the organ model with a Gaussian process. Finally, a high-quality image segmentation result can be automatically achieved using Bayes RW algorithm. Comparing our method with conventional RW and state-of-the-art interactive segmentation methods, our results show an improvement in the accuracy for liver segmentation ($p < 0.001$).

1. Introduction

Segmentation of organ from CT volume is an important prerequisite for computer-aided surgery, computer-assisted intervention, and image-guided surgery. The accurate segmentation of the organ from clinical CT images is considered a challenging task: Large variations in shape make an accurate segmentation difficult, and existing lesions (e.g., tumors) exhibit considerable variation for the organ anatomical structure. To accurately segment an organ, various approaches have been proposed in literatures [1–8], such as intensity-based [9–11], classification-based [12, 13], clustering-based [14–18], statistical shape model- (SSM-) based [19, 20], probabilistic atlas- (PA-) based [21–25], active contour- (AC-) based [26, 27], and watershed-based [28, 29] segmentation methods. However, the main challenge of the abovementioned methods is the fast and efficient segmentation of large image data. This can be observed particularly in medical applications where a resolution of three-dimensional CT and MRI body scans constantly increases.

Recently, a growing interest is attracted by an interactive graph-based image segmentation algorithms such as graph cut (GC) [30–36] and random walker (RW) [37–41] algorithms. The random walker algorithm represents a recent noteworthy development in the weighted graph-based interactive segmentation methods. This technique with user interaction is more suitable for volumetric medical images to guarantee the reliability, accuracy, and fast speed demands.

However, due to the classical RW algorithm definitions on the weighted graphs, for a high-resolution volumetric medical image, RW method needs to construct the corresponding large-scale graph to solve the resulting sparse linear system, which leads to high computation cost: the long computation time and the high memory usage. Hence, over the past years, a large amount of research has been conducted to extend and enhance the random walker algorithm. Grady et al. [40] extended the classical RW segmentation approach by combining the regional intensity priors. The sparse linear equations can be addressed by the preconditioned conjugate gradient to achieve an acceptable memory consumption and easy parallelization. In [41], the computational demands with RW are alleviated by introducing an "offline" precomputation before user interaction with RW in real-time "online." Using a similar principle, an offline precomputation was used to further speed up the online segmentation in [42]. Both methods used the "offline" and "online" strategies to minimize the time spent waiting. In addition, Goclawski et al. [43] proposed a superpixel-based random walker method to reduce the graph size, while the computation time increases linearly with the number of superpixels. The accuracy of superpixels plays an immediate decisive role in the process of organ segmentation.

To resolve these limitations, in our previous research [44], we proposed a knowledge-based segmentation framework for the volumetric medical image in a slice-by-slice manner based on the classical random walker. This algorithm employs the previous segmented slice as the prior knowledge for automatically setting the object/background seed points for the adjacent slices. It can reduce the graph scale and significantly speed up the optimization procedure of the graph. However, the classical RW algorithm was designed to be a general purpose interactive segmentation method, such that a user could mark a few pixels with an arbitrary number of labels and expect a quality result, regardless of the data set or the segmentation goal. Segmentation of a medical image ignores itself absolute intensity and shape information. If a consistent intensity and shape profile characterize an object of interest, then this information should be incorporated into the RW segmentation process.

Taking these into consideration, in our study, we extended a classical random walker algorithm by incorporating the prior (shape and intensity) knowledge in the optimization of sparse linear system. The objective of our work is to combine the prior knowledge with the spatial cohesion of the random walker algorithm in a principled way that produces the correct result. Based on the extended random walker, we applied a knowledge-based segmentation framework for the volumetric medical image in a slice-by-slice manner. Our strategy is to employ the previous segmented slice to obtain the prior (shape and intensity) knowledge of the target organ for the adjacent slice. With a small number of user-defined seed points, we can obtain the segmentation results of the start slice in the volume which can be used as the prior knowledge of the target organ. According to this prior knowledge, the object/background seed points are automatically defined and the corresponding Bayes model can be generated. Integrating this Bayes model into the RW sparse system, the organ is automatically segmented for the adjacent slice.

The remainder of this paper is organized as follows. Section 2 presents a brief recapitulation of the random walker algorithm and then extends to incorporate the prior (shape and intensity) knowledge. Section 3 elaborates our proposed knowledge-based framework using the extended RW with the Bayes model. Section 4 contains experimental work, and Section 5 discusses the implementation of our method, followed by the conclusion (Section 6).

2. Development

The random walk algorithm treats image segmentation as an optimization problem on a weighted graph, where each node represents a pixel or voxel. Therefore, we firstly define the graph that we are working on. We use the following notations for the rest of the paper. Given an image, I, a graph consists of $G = (V, E)$ with vertices (nodes) $v \in V$ and edges $e \in E$. Each node v_i in V uniquely identifies an image pixel x_i. An edge, e, spanning two vertices v_i and v_j, is denoted by e_{ij}. A weighted graph assigns a weight to each edge. The weight of an edge, e_{ij}, is denoted by w_{ij}. It represents the similarity between two neighboring nodes v_i and v_j. The degree of a vertex is $d_i = \sum_j w_{ij}$ for all edges e_{ij} incident on v_i.

2.1. Review of Random Walker Method. The random walker segmentation algorithm of [37] computes the probability, for each pixel, that a random walker leaving that pixel will first arrive at a foreground seed before arriving at a background seed. It was shown in [37] that these probabilities may be calculated analytically by solving a linear system of equations with the graph Laplacian matrix. The Laplacian matrix is defined as

$$L_{ij} = \begin{cases} d_i & i = j \\ -w_{ij} & v_i \text{ and } v_j \text{ are adjacent nodes} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where L_{ij} is indexed by vertices v_i and v_j. $w_{ij} = \exp(-\beta (I_i - I_j)^2)$ is the edge weight, and I_i and I_j indicate the image intensity at vertices v_i and v_j, respectively. β represents a tuning constant that depends on the user.

Given a weighted graph, a set of marked (labeled) nodes, V_M, and a set of unmarked nodes, V_U, such that $V_M \cup V_U = V$ and $V_M \cap V_U = \varnothing$, we would like to label each node $v_i \in V_U$ with a label s. $s = 1$ stands for the foreground, and $s = 2$ stands for the background. Assuming that each node $v_j \in V_M$ has also been assigned with a label s, we can compute the probabilities, x_i^s, that a random walker leaving node v_i arrives at a marked node v_j by solving the minimization of

$$E^s_{\text{internal}} = \frac{1}{2} x^{sT} L x^s. \tag{2}$$

All nodes V are divided into two sets: the marked (prelabeled) nodes V_M and unlabeled (i.e., free) nodes V_U. Therefore, the above function can be reformulated as follows:

$$E^s_{\text{internal}} = \frac{1}{2} \left[x_M^{sT} x_U^{sT} \right] \begin{bmatrix} L_M & B \\ B^T & L_U \end{bmatrix} \begin{bmatrix} x_M^s \\ x_U^s \end{bmatrix}. \tag{3}$$

Minimization of (3) with respect to x_U^s, the random walker problem can be solved by the following system of equations:

$$L_U x_U^s = -B^T x_M^s. \tag{4}$$

The variable x_U^s represents the set of probabilities corresponding to unmarked nodes; x_M^s is the set of probabilities corresponding to marked nodes (i.e., "1" for foreground nodes and "0" for background nodes). By virtue of x_i being a probability,

$$\sum_{s=1}^{2} x_i^s = 1 \quad \forall i. \tag{5}$$

The random walk algorithm is explained in detail elsewhere [37]. Next, we will now present how the incorporation of the Bayes model into the above framework yields a segmentation algorithm.

2.2. Random Walker with Bayes Model. According to the above priori knowledge, we can calculate a posterior probability $p(s|I_i)$ at the node v_i which belongs to the label s. Assuming that each label is equally likely, Bayes theorem gives the probability that a node v_i belongs to label s as

$$x_i^s = \frac{p(s|I_i)}{\sum_{q=1}^{2} p(q|I_i)} = \frac{p(I_i|s) \times p(s)}{\sum_{q=1}^{2} p(I_i|q) \times p(q)}, \tag{6}$$

where $p(I|s)$ is the likelihood map for an organ and $p(s)$ is the shape map for the targeted organ. $p(s)$ can be obtained by dilating the targeted organ region in the previous segmented slice. $p(I|s)$ can be estimated by the previous segmented slice of the organ. $s = 1$ is the foreground, and $s = 2$ is the background.

Equation (6) can be also written in vector notation:

$$\left(\sum_{q=1}^{2} \Lambda^q\right) x^s = p(s|I), \tag{7}$$

where Λ^s is a diagonal matrix with the values of $p(s|I)$ on the diagonal.

According to (6), the minimum energy distribution for the external function is

$$E_{\text{external}}^s = \sum_{q=1, q\neq s}^{2} x^{qT} \Lambda^q x^q + (x^s - 1)^T \Lambda^s (x^s - 1). \tag{8}$$

To incorporate the posteriori probability function (external term) into the RW algorithm (internal term), we may optimize the following energy:

$$E_{\text{total}}^s = E_{\text{internal}}^s + \gamma E_{\text{external}}^s. \tag{9}$$

The first term is the driving force behind the spatial cohesion of the random walker algorithm. The second term is a Bayes penalty term with the weight γ used to guarantee robustness to small disconnected pieces. The used Bayes model is generated according to the prior knowledge of an organ: shape and intensity. In this work, we set the weight to $\gamma = 0.01$.

The minimum energy of the above equation is obtained when x_s satisfies the solution to

$$\left(L + \gamma \sum_{q=1}^{2} \Lambda^q\right) x^s = \gamma p(s|I). \tag{10}$$

Optimizing this energy leads to the system of linear equations:

$$\left(L_U + \gamma \sum_{q=1}^{2} \Lambda^q\right) x_U^s = \gamma p(s|I_U) - B^T x_M. \tag{11}$$

The usage of the proposed Bayes-based RW algorithm is strongly limited by the enormous size of the graph represented in 3D volumetric medical image and the necessity of solving a huge sparse linear system. It results in the relative increase of the unlabeled seed points relative to a 2D image. Hence, in order to estimate the probability of each unlabeled seed point, the extended RW algorithm needs to calculate the larger inverse matrix $\left(L_U + \gamma \sum_{q=1}^{2} \Lambda^q\right)^{-1}$, which leads to high computation costs: long computation time and high memory usage. We integrated our extended RW algorithm into a knowledge-based framework to make it more suitable and workable for our application. The following details our knowledge-based framework and results.

3. Knowledge-Based Framework

Our knowledge-based strategy employs the previous segmented slice as the prior (shape and intensity) knowledge of the target organ for automatic segmentation of the adjacent slice. Using a small number of user-defined seed points, we can obtain the segmentation results of the start slice of the volume for use as the prior knowledge of the target organ. According to the prior knowledge, the object/background seed points can be dynamically updated for the adjacent slice by combining the narrow band threshold (NBT) method and the organ model with a Gaussian process. Meanwhile, the corresponding Bayes model can be generated. Finally, an extended Bayes-based random walker algorithm is applied to automatically segment the whole volume in a slice-by-slice manner. In our work, "object" means the target organ to be segmented and "background" means the other tissues except the target organ. The whole procedure of the proposed approach is shown in Figure 1. In this method, there is a three-step pipeline consisting of the following:

(1) Selecting and segmenting the start slice, as shown in the middle-part of Figure 1: (a) Manually defining the object/background seed points. (b) Generating a Gaussian model (GM) using the seed points. (c)

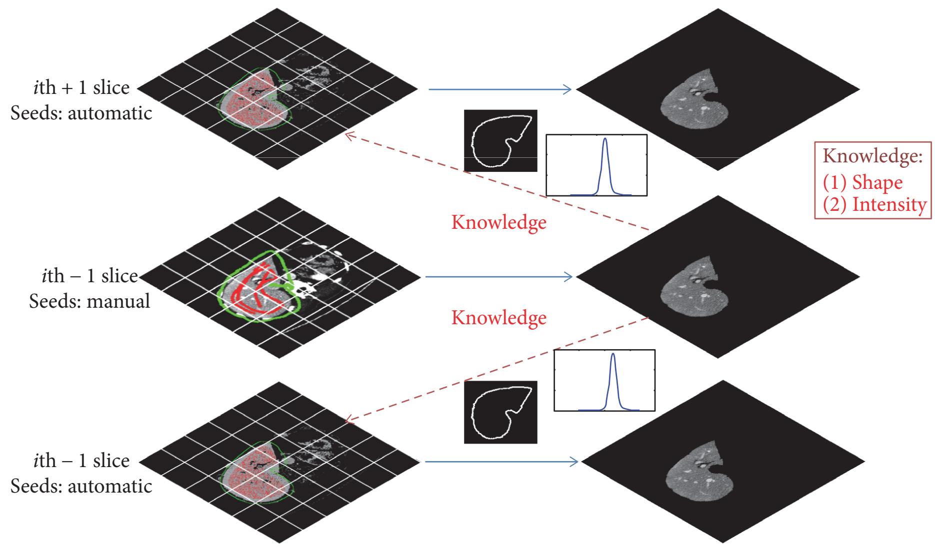


Figure 1: The whole procedure of our knowledge-based method.

Segmenting the organ ("Candicate Pixels" for the liver) using the classical RW method.

(2) Segmenting the adjacent slice, as shown in the upper-part and bottom-part of Figure 1: (a) Generating a Gaussian model (GM) according to the previous segmented organ (intensity knowledge). (b) Automatic setting the object/background seeds based on the restricted region by morphological operation of the previous segmented organ (shape knowledge). (c) Refining the seed points based on NBT. (d) Segmenting the organ using our proposed Bayes-based RW methods. Thus, it automatically segments the whole organ in the remaining slices based on the updated prior knowledge of the organ.

(3) Smoothing the boundary of the whole volume: Finally, the boundary of the output volume is smoothed by "Fourier transform" that forms the final organ surface.

In the following section, we will introduce the start slice segmentation, the GM generation, and automatic seed point selection which integrate the prior intensity and shape knowledge of the previous segmented organ.

3.1. Interactive Segmentation of the Start Slice. Our proposed segmentation is a slice-by-slice method. There are two main steps in our proposed method. The first step is to segment the start slices interactively, and the second step is to segment other remaining slices automatically based on the segmented start slices. The aim of the first step (interactive segmentation of the start slices) is to find the initial region of the target organ (liver) so that it can be used as prior (intensity and shape) knowledge of the organ as the following steps for automatic segmentation.

The process of the first interactive segmentation of two start slices is shown in Figure 2 and involves four steps: (1) manually select one axial start slice. Scanning an input CT volume along the axial axis to find one slice in which the organ has the relative larger cross section in the axial plane; (2) manually define the object/background seeds on this start slice; (3) automatically generate the thresholded images based on the constructed Gaussian model (GM) using these seeds. To remove the intercostal muscles and the other nonobject parts, the object seeds are employed to construct the approximate intensity models for this organ using the Gaussian model (GM). After estimating the statistical intensity model, the constructed model is thresholded to find "Candidate Pixels" for the organ; (4) automatically segment the thresholded images. The final step of this process is to segment the thresholded image based on "Candidate Pixels" by the classical RW method.

3.2. Automatic Segmentation of the Adjacent Slice

3.2.1. GM for Generation of the Thresholded Image. Constructing a Gaussian model (GM) [45] is aimed to estimate a new preprocessed image of the target organ so that it can more easily distinguish the difference between the target organ and other tissues. As explained in the last section, the initial segmented slice can be used to estimate the statistical parameters of the liver model for the current slice. Due to the existence of a large number of the liver pixels, estimation of the statistical parameters can be trusted. A Gaussian model is employed to estimate the intensity distribution of the liver. The Gaussian model is given by

$$p(I_i|s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(\frac{-(I_i-\mu_s)^2}{2\sigma_s^2}\right),$$
$$p(I_i|s) = \frac{p(I_i|s)}{\sum_{q=1}^{2} p(I_i|q)}, \tag{12}$$

where the parameters mean μ_s and variance σ_s^2 can be estimated by the marked seed points or the previous segmented

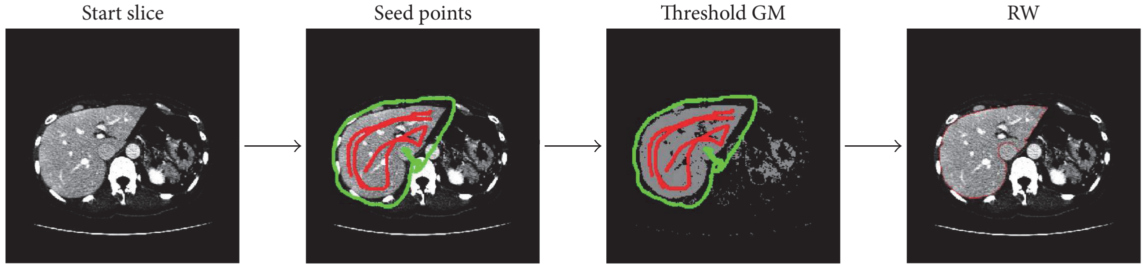


Figure 2: Interactive segmentation of the start slice in CT image.

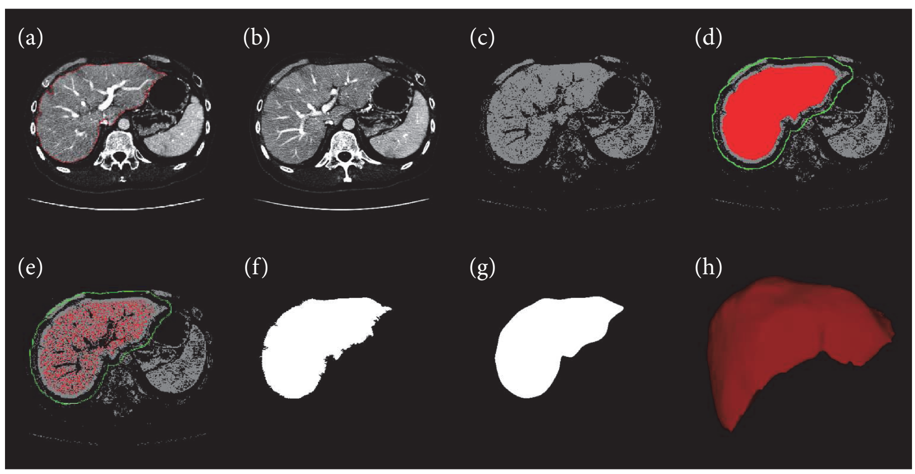


Figure 3: Steps of the RWBayes method. (a) The segmented liver (red) of the previous slice; (b) the current slice; (c) candidate pixel by thresholding the GM; (d) the rough object (red) and background (green) seed points; (e) the fine seed points using a NBT method; (f) the initial segmentation result by RWBayes; (g) smoothing the boundary by Fourier transform; and (h) visualisation of the segmented liver volume.

slice of the organ. I_i indicates the image intensity at the node v_i. $s = 1$ is the object, and $s = 2$ is the background.

The intensity models are automatically determined for each slice according to the segmented organ in the previous slice. Furthermore, in order to remove some nonobject parts and obtain an accurate result, we threshold the output of this intensity model by discarding probabilities less than 0.5, so it can generate a likelihood map of the object. Comparison of the original CT image (Figure 3(b)) with the corresponding intensity model (Figure 3(c)) revealed that the liver can be more easily distinguished from other tissues. However, for the background, the likelihood map keeps the original probability value without thresholding.

3.2.2. Automatic Setting of Seed Points. The main assumption in our method is that it can determine the approximate prior (shape and intensity) knowledge for the organ. Due to a slice-by-slice technique that is applied to segment the organ in our method, the user segments one slice in the volume to define this prior knowledge, and consequently, they are automatically updated for the nearby slices. In this approach, assuming the consequent slices of the same patient have a high correlation, the boundary of the organ in the next slice does not go far from its border in the previous slice. Thus, a defined shape constraints based on the previous slice can be used to roughly select the object/background seed points for the adjacent slice.

Assuming the cross-section of the liver in the ith slice is divided into m parts and the region of the organ for each part ($\mathrm{Mask}_{i,j}$, $1 \le j \le m$) is known, corresponding to the part j in the $(i+1)$th slice, the object and background seeds can be defined by the following equation:

$$\begin{aligned} \mathrm{BS}_{i+1,j} &= \left(\mathrm{Mask}_{i,j} \oplus \mathrm{BE}_{\mathrm{Dilation2}}\right) - \left(\mathrm{Mask}_{i,j} \oplus \mathrm{BE}_{\mathrm{Dilation1}}\right), \\ \mathrm{FS}_{i+1,j} &= \mathrm{Mask}_{i,j} \oplus \mathrm{RE}_{\mathrm{Erosion}}, \end{aligned} \tag{13}$$

where $\mathrm{Mask}_{i,j}$ is the mask of the organ corresponding to the jth part in the slice i. $\mathrm{BE}_{\mathrm{Dilation1}}$ and $\mathrm{BE}_{\mathrm{Dilation2}}$ are the structuring elements used for dilation in the region. $\mathrm{RE}_{\mathrm{Erosion}}$ is the structuring elements used for erosion in the region. These elements are empirically selected to be disks with a radius of $\mathrm{BE}_{\mathrm{Dilation1}} = 10$ pixels, $\mathrm{BE}_{\mathrm{Dilation2}} = 8$ pixels, and $\mathrm{BE}_{\mathrm{Erosion}} = 8$ pixels.

The background seed points are directly selected in the current slice in the region $\mathrm{BS}_{i+1,j}$ which can be considered as accurately seeded points outside the liver's boundary. However, as shown in Figure 3(d), it can be seen that there were still a lot of false positives (other tissues) in the $\mathrm{FS}_{i+1,j}$ despite eroding the liver region for the previous slice, because we cannot segment the previous slice accurately and there still exists a variation of liver shape for different slices.

3.2.3. Refinement of Seed Points. As already explained above, it can dynamically update the parameters of GM model for the following slices. If the intensity model of the liver includes the parameters μ and σ, we can threshold this component in the narrow region $[\mathrm{TL}, \mathrm{TH}]$ to find the fine seed points corresponding to the candidate liver pixel.

$$\begin{aligned} \mathrm{TL} &= \mu - \beta\sigma, \\ \mathrm{TH} &= \mu + \beta\sigma. \end{aligned} \tag{14}$$

We empirically found that the values of β are in the range $[0.05, 0.3]$ corresponding to low-contrast and high-contrast datasets.

In addition, it can been seen from Figure 3(d), since the defined region $\mathrm{FS}_{i+1,j}$ may include the nonliver part (such as vessels); we can threshold the narrow band to achieve more accurate object seeds (Figure 3(e)). Thus, for a pixel located in the region $\mathrm{FS}_{i+1,j}$, if the intensity value of this pixel belongs to the narrow range $[\mathrm{TL}, \mathrm{TH}]$, it is considered as an object seed. After estimating the "Candidate Pixels" and the fine object/background seeds for the current slice, the Bayes-based RW algorithm is applied to segment the liver (Figure 3(f)).

3.3. Smoothing the Boundary of the Whole Volume. However, the boundary of the segmented object obtained in the last step is not smooth, as shown in Figure 3(f). If the coordinates of the boundary points are analyzed by the Fourier transform (FT), they contain a significant number of high-frequency components. According to the definition of the FT, the coordinates (x, y) are transformed from the spatial domain into the frequency domain as

$$\begin{aligned} F_x(k) &= \sum_{j=1}^{N} x(j) e^{\left(-\frac{2\pi i}{N}(j-1)(k-1)\right)} \\ F_y(k) &= \sum_{j=1}^{N} y(j) e^{\left(-\frac{2\pi i}{N}(j-1)(k-1)\right)} \end{aligned} \quad i = \sqrt{-1}, \tag{15}$$

where N is the number of the boundary points that are usually greater than 100. The boundary is smoothed by removing the high-frequency components, while the useful (information bearing) low-frequency components are retained. Hence, the first 15 components in frequency domain are kept and then transferred into the spatial domain (Figure 3(g)).

$$\begin{aligned} x(j) &= \sum_{k=1}^{15} F_x(k) e^{\left(\frac{2\pi i}{N}(j-1)(k-1)\right)} \\ y(j) &= \sum_{k=1}^{15} F_y(k) e^{\left(\frac{2\pi i}{N}(j-1)(k-1)\right)} \end{aligned} \quad i = \sqrt{-1}. \tag{16}$$

4. Results

4.1. Database. Our dataset included 26 CT images of the abdominal region with a resolution of $0.683 \times 0.683 \times 1\,\mathrm{mm}^3$ and a size of $512 \times 512 \times (159\text{–}263)$ pixels. All of the data were stored in DICOM image format with a depth of 12 bits per pixel. These data were acquired by GE LightSpeed Ultra scanners with eight detectors. The large variation of liver images was an important feature in the evaluation of our segmentation method. Hence, data were acquired from normal and pathological cases between 20 and 75 years old. The sample contained 20 normal cases and 6 pathological cases: no. 1 to no. 20 were normal cases and no. 21 to no. 26 belonged to pathological cases. Therein, patients (pathological cases) were those who were suspected of having a disease, such as chronic liver disease, and were scanned in the course of diagnosis. In order to make a quantitative evaluation for our proposed method, the liver was segmented for each image (i.e., subject) manually as the ground truth. The segmentation was performed under the guidance of a physician in order to obtain accurate liver volumes. This study was conducted with the approval of the institutional review boards at University Ethics Committee, and all data provided written informed consent.

The proposed algorithm was implemented in a MC-OS-based personal computer (Intel®Corei7 2.5GHz and 16GB-DRAM). The programming environment was coded in the MATLAB environment. Visualization of the shapes was performed using VTK [46] in C++ languages.

4.2. Quantitative Measurement. To measure the accuracy of our method, we compared it with the conventional RW method and the state-of-the-art interactive segmentation algorithms by two metrics.

4.2.1. Dice Coefficient (Dice). The dice coefficient is one of the most popular methods to evaluate segmentation accuracy. This metric is given in percent and based on the voxels of two binary 3D volumes, with V_{manual} as the manually and V_{auto} as the automatically segmented organs.

$$\mathrm{Dice} = \frac{2|V_{\mathrm{manual}} \cap V_{\mathrm{auto}}|}{|V_{\mathrm{manual}}| + |V_{\mathrm{auto}}|} \times 100\%. \tag{17}$$

4.2.2. Volumetric Overlap Error (VOE). The volumetric overlap error between two sets of voxels V_{manual} and V_{auto} is given in percent. This ratio is also known as Tanimoto or Jaccard coefficient.

$$\mathrm{VOE} = \frac{|V_{\mathrm{manual}} \cap V_{\mathrm{auto}}|}{|V_{\mathrm{manual}} \cup V_{\mathrm{auto}}|} \times 100\%. \tag{18}$$

4.3. Quantitative Validation of Liver Segmentation. To investigate the performance of our proposed segmentation method, we applied our proposed RWBayes method to 26 clinical CT volumes which are described in the previous section. The segmentation results of two typical cases are shown in Figure 4. The results in Figure 4 proved that performing the RWBayes method to segment the livers can give us accurate results. A common difficulty for computer-aided

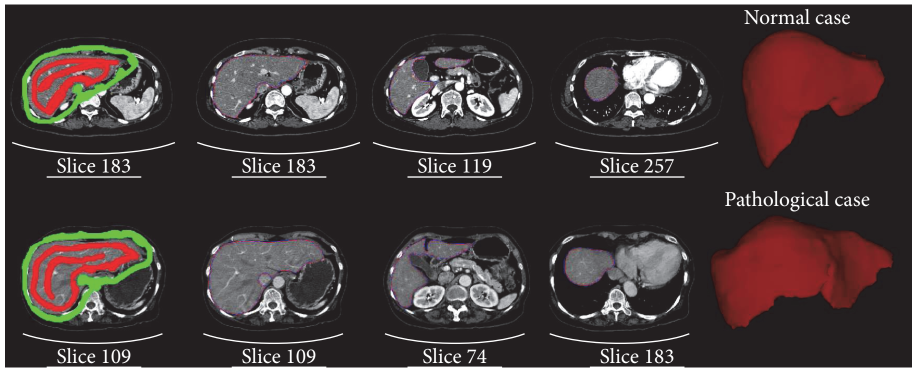


FIGURE 4: Comparison of the manual segmentation (blue) with the segmentation results of our method (red). The first row is the segmentation result in case 9. The second row is the segmentation result of pathological case with the unusual liver shape in case 22.

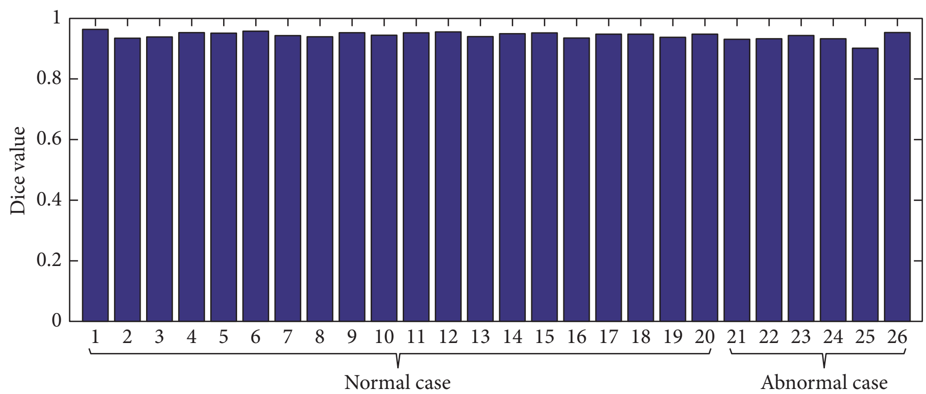


FIGURE 5: Our technique performed on 26 CT scans with Dice measurement. The first 20 data points are normal cases, and the remaining 6 data points are pathological cases.

liver segmentation is the erroneous inclusion of heart volumes, which our method robustly avoided. It confirmed the ability of our method to segment the livers with a precision segmentation result.

Additional challenges come from enlarged livers, where the liver has large shape variations which made it very difficult to be segmented. Taking this limitation into consideration, in this research, our technique performed on 26 CT scans that combined normal cases and pathological cases with large morphological variations. Figure 4 shows the liver segmentation result from one pathological case. It proved the performance of our proposed algorithm which was robust for segmenting the liver in the pathological cases with large morphological variations.

Apart from a visual inspection, a quantitative evaluation was conducted. Figure 5 gave a more clear depiction of the corresponding accurate results of 26 cases. The first 20 data points correspond to normal cases (the average Dice is 0.946), and the remaining 6 data points are pathological cases (the average Dice is 0.930). Regarding the result of applying our method to synthetic shapes, we can conclude that our proposed method was robust in addressing the segmentation of the liver (with the average Dice's similarity coefficient = 0.942). Future research directions will include applying our method on more datasets in order to more accurately evaluate the performance.

4.4. Qualitative Comparison of Interactive Segmentation Methods. To evaluate the effectiveness of the proposed method (RWBayes), RWBayes was compared with the classical random walk (RW3D) [37]. Considering the memory usage demands for applying the RW3D algorithm to the computer, we resized all of our datasets ($512 \times 512 \times (159–263)$ pixels) into the size of $128 \times 128 \times 36$ pixels. Moreover, we also compared our proposed method with a knowledge-based framework using the classical random walker and narrow band threshold (RWNBT) [44], in which the RWNBT did not generate a thresholded image based on the constructed Gaussian mixture model according to the previous segmented liver.

Quantitative and comparative results from applying the RW3D, RWNBT, and RWBayes methods for the liver segmentation are presented in Figure 6. In order to intuitively make a comparison between our proposed RWBayes and RW3D methods, it was unreasonable to give only one start slice with the corresponding segmentation result. It was necessary to show different slices for one data corresponding to a point on the curve with $128 \times 86 \times 33$ pixels. The

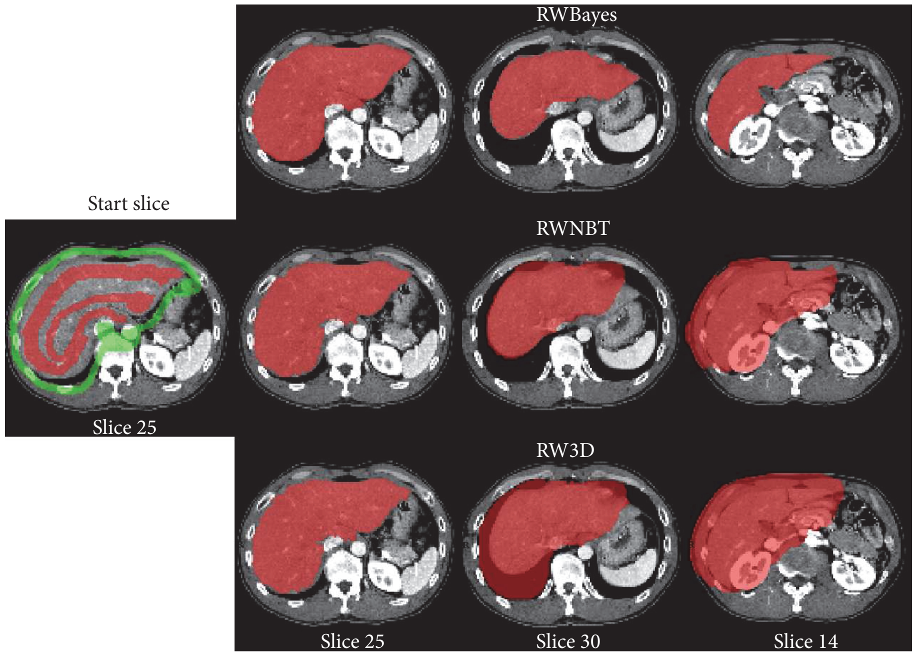


Figure 6: Comparison of the liver segmentation results with RWBayes method, RWNBT method, and RW3D method in case 6.

Table 1: Segmentation accuracy obtained by the state-of-the-art methods for the liver on 26 CT scans.

	RW3D [37]	GC [34]	IKM [14]	RWNBT [44]	RWBayes
Dice	0.573	0.857	0.894	0.687	0.934
VOE	0.404	0.758	0.810	0.526	0.874
Runtime (sec)	45.800	1.828	2.530	1.781	1.231

red images were segmented liver slices, which were overlaid with the original CT slices. The simulation verifies that the performance of RWBayes was significantly better than the RW3D and RWNBT methods for segmenting the liver.

In order to make a comparison with the state-of-the-art interactive segmentation algorithms, we also compared the results using the graph cut algorithm (GC) [34] and interactive K-means algorithm (IKM) [14]. Table 1 clearly depicts the merits of our method by listing the comparative results with the average of Dice, VOE, and runtime between automated and manual segmentations for all 26 test CT scans. Computation time is an important metric for evaluating one segmentation algorithm. For the classical RW algorithm, the basis of RW method is a large, sparsely occupied linear equations, whose size corresponds to the number of voxels in the 3D image. Hence, it exhibited slowness for solving 3D image segmentation. A significant reduction in runtime values using RWBayes-based segmentation compared with those based on RW3D was confirmed. Meanwhile, the accuracy of RWBayes was observed to have significantly higher Dice/VOE than the state-of-the-art interactive segmentation methods. To directly demonstrate the performance of our proposed method, in respect to the statistical significance analysis, the p value was the probability of obtaining a test statistic result that was actually observed. These statistical tests demonstrated that our proposed RWBayes approach yields the high precision results with respect to the conventional RW3D method ($p < 0.001$).

5. Discussion

This paper introduced a new knowledge-based framework for the organ segmentation using the RWBayes method. The proposed method segmented an organ based on a set of prior knowledge. Prior knowledge included the approximate shape of an organ (shape knowledge) and statistical parameters of the organ's intensities (intensity knowledge). According to a prior knowledge of an organ, the proper selection of object/background seeds was performed skillfully for our method to accurately segment the organ from the CT image.

The basic idea of the proposed method is based on the high correlation between adjacent slices. Seed points for the current slice are automatically generated according to the prior knowledge from the segmented organ region of the previous slice. As shown in Figure 5, precision results were achieved in our experiments as we used high-resolution data.

In practical clinics, however, CT images exist in various resolutions. In general, thin slices (high resolution) correspond to strong correlation while thick slices (low resolution) correspond to weak correlation.

In order to verify the effect of resolutions on our RWBayes method, a typical CT image (a resolution of $0.683 \times 0.683 \times 1.25\,\mathrm{mm}^3$ and a size of $512 \times 512 \times 159$

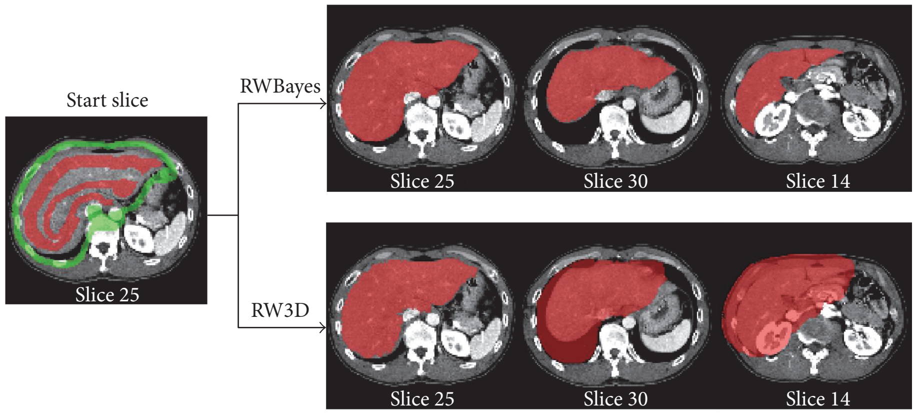


Figure 7: Effect of resolution on segmentation accuracy for case 1.

pixels) is resized into 7 different resolutions in the axial-axis (z-axis) and then segmented with the same seed points. Figure 7 indicates that our proposed technique can be performed on the CT scans with large resolution variations. Regardless of image resolution, satisfactory segmentation results were achieved. In conclusion, our RWBayes method was robust in segmenting the livers from CT images of various resolutions. The simulation results prove the high capacity of our proposed RWBayes method for the organ segmentation using various resolutions of CT scans.

6. Conclusion

In this paper, we proposed a novel knowledge-based framework for organ segmentation using the RWBayes algorithm. A prior knowledge of the previous segmented organ was integrated into our strategy and has the following benefits: (1) small-scale graph; (2) automation of object/background seed setting according to the prior knowledge of the already segmented slices; and (3) robust segmentation technique by combing a Bayes model of an organ into the sparse system to calculate the probability of each unmarked node. The evaluation of the results demonstrated the high precision of the proposed approach. Compared with the conventional RW and the state-of-the-art interactive segmentation methods, our proposed method can significantly improve the segmentation accuracy ($p < 0.001$). As for future applications, the proposed method can be extended to segment other organs.

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 Army Research office under Award no. W911NF-15-1-0521 from the USA and in part by the MEXT Support Program for the Strategic Research Foundation at Private Universities (2013–2017) from Japan.

References

[1] S. Umetsu, A. Shimizu, H. Watanabe, H. Kobatake, and S. Nawano, “An automated segmentation algorithm for CT volumes of livers with atypical shapes and large pathological lesions,” *Journal of IEICE Transactions on Information and Systems*, vol. E97-D, no. 4, pp. 951–963, 2014.

[2] P. Karasev, I. Kolesov, K. Fritscher, P. Vela, P. Mitchell, and A. Tannenbaum, “Interactive medical image segmentation using PDE control of active contours,” *IEEE Transactions on Medical Imaging*, vol. 32, no. 11, pp. 2127–2139, 2013.

[3] D. Mahapatra, “Analyzing training information from random forests for improved image segmentation,” *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1504–1512, 2014.

[4] J. L. Peng, Y. Wang, and D. X. Kong, “Liver segmentation with constrained convex variational model,” *Pattern Recognition Letters*, vol. 43, pp. 81–88, 2014.

[5] L. DF, Y. Wu, G. Harris, and W. L. Cai, “Iterative mesh transformation for 3D segmentation of livers with cancers in CT images,” *Computerized Medical Imaging and Graphics*, vol. 43, pp. 1–14, 2015.

[6] A. Shimizu, R. Ohno, T. Ikegami, H. Kobatake, S. Nawano, and D. Smutek, “Segmentation of multiple organs in non-contrast 3D abdominal CT images,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 2, no. 3-4, pp. 135–142, 2007.

[7] W. Xiong, S. H. Ong, Q. Tian et al., “Construction of a linear unbiased diffeomorphic probabilistic liver atlas from CT images,” in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 1773–1776, Cairo, Egypt, November 2009.

[8] A. H. Foruzan and Y. W. Chen, “Improved segmentation of low-contrast lesions using sigmoid edge model,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 7, pp. 1267–1283, 2016.

[9] X. L. Zhang, X. F. Li, and Y. C. Feng, “A medical image segmentation algorithm based on bi-directional region growing,” *Optik - International Journal for Light and Electron Optics*, vol. 126, no. 20, pp. 2398–2404, 2015.

[10] Z. G. Pan and J. F. Lu, “A Bayes-based region-growing algorithm for medical image segmentation,” *Journal of Computing in Science and Engineering.*, vol. 9, no. 4, pp. 32–38, 2007.

[11] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[12] E. A. Rikxoort, Y. Arzhaeva, and B. Ginneken, "Automatic segmentation of the liver in computed tomography scans with voxel classification and atlas matching. MICCAI workshop 3-D segmental," *Clinic: A Grand Challenge*, pp. 101–108, 2007.

[13] K. Kasiri, K. Kazemi1, M. J. Dehghani, and M. S. Helfroush, "Atlas-based segmentation of brain MR images using least square support vector machines," in *2010 2nd International Conference on Image Processing Theory, Tools and Applications (IPTA 2010)*, pp. 306–310, Paris, France, July 2010.

[14] A. H. Foruzan, Y. W. Chen, R. A. Zoroofi et al., "Segmentation of liver in low-contrast images using K-means clustering and geodesic active contour algorithms," *IEICE Transactions on Information and Systems*, vol. E96-D, pp. 798–807, 2013.

[15] Y. W. Chen, K. Tsubokawa, and A. H. Foruzan, "Liver segmentation from low contrast open MR scans using K-means clustering and graph-cuts," in *Advances in Neural Networks - ISNN 2010*, L. Zhang, B. L. Lu and J. Kwok, Eds., vol. 6064 of Lecture Notes in Computer Science, pp. 162–169, Springer, Berlin, Heidelberg, 2010.

[16] B. N. Li, C. K. Chui, S. Chang, and S. H. Ong, "Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation," *Computers in Biology and Medicine*, vol. 41, no. 1, pp. 1–10, 2011.

[17] Y. He, M. Y. Hussaini, J. Ma, B. Shafei, and G. Steidl, "A new fuzzy c-means method with total variation regularization for segmentation of images with noisy and incomplete data," *Pattern Recognition*, vol. 45, no. 9, pp. 3463–3471, 2012.

[18] Z. X. Ji, Q. S. Sun, and D. S. Xia, "A modified possibilistic fuzzy c-means clustering algorithm for bias field estimation and segmentation of brain MR image," *Computerized Medical Imaging and Graphics*, vol. 35, no. 5, pp. 383–397, 2011.

[19] X. Zhang, T. Jie, K. X. Deng, Y. F. Wu, and X. L. Li, "Automatic liver segmentation using a statistical shape model with optimal surface detection," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, Part 2, pp. 2622–2626, 2010.

[20] T. Heimann and H. P. Meinzer, "Statistical shape models for 3D medical image segmentation: a review," *Journal of Medical Image Analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[21] H. Lamecker, T. Lange, and M. Seebaee, *Segmentation of the Liver Using a 3D Statistical Shape Model*, ZIB Technology Report, Zuse Institute, Berlin, German, 2004.

[22] T. Okada, R. Shimada, M. Hori et al., "Automated segmentation of the liver from 3D CT images using probabilistic atlas and multi-level statistical shape model," *Academic Radiology*, vol. 15, no. 11, pp. 1390–1403, 2008.

[23] M. G. Linguraru, J. K. Sandberg, Z. Li, F. Shah, and R. M. Summers, "Automated segmentation and quantification of liver and spleen from CT images using normalized probabilistic atlases and enhancement estimation," *International Journal of Medical Physics*, vol. 37, no. 2, pp. 771–783, 2010.

[24] C. H. Dong, Y. W. Chen, A. H. Foruzan et al., "Segmentation of liver and spleen based on computational anatomy models," *Computers in Biology and Medicine*, vol. 67, pp. 146–160, 2015.

[25] H. Park, P. H. Bland, and C. R. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *IEEE Transactions on Medical Imaging*, vol. 22, no. 4, pp. 483–492, 2003.

[26] M. Esfandiarkhani and A. H. Foruzan, "A generalized active shape model for segmentation of liver in low-contrast CT volumes," *Computers in Biology and Medicine*, vol. 82, pp. 59–70, 2017.

[27] V. Caselles, F. Catte, T. Coll, and F. Dibos, "A geometric model for active contours in image processing," *Numerische Mathematik*, vol. 66, no. 1, pp. 1–31, 1993.

[28] V. Grau, A. U. Mewes, M. Alcaniz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 447–458, 2004.

[29] N. Salman and C. Q. Liu, "Image segmentation and edge detection based on watershed techniques," *International Journal of Computers and Applications*, vol. 25, no. 4, pp. 258–263, 2003.

[30] Y. Boykov, "Graph cuts and efficient N-D image segmentation," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.

[31] Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proceedings Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, vol. 1, pp. 105–112, Vancouver, BC, Canada, July 2001.

[32] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[33] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[34] A. Afifi and T. Nakaguchi, "Liver segmentation approach using graph cuts and iteratively estimated shape and intensity constraints," in *IEEE International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2012)*, vol. 7511 of *Lecture Notes in Computer Science*, pp. 396–403, Springer, Berlin, Heidelberg, 2012.

[35] A. K. Sinop and L. Grady, "A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm," in *2007 IEEE 11th International Conference on Computer Vision (ICCV 2007)*, pp. 1–8, Rio de Janeiro, Brazil, October 2007.

[36] W. Casaca, L. G. Nonato, and G. Taubin, "Laplacian coordinates for seeded image segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, pp. 384–391, Columbus, OH, USA, June 2014.

[37] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006.

[38] L. Grady, T. Schiwietz, S. Aharon, and R. Westermann, "Random walks for interactive organ segmentation in two and three dimensions: implementation and validation," in *IEEE International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2005)*, vol. 3750, pp. 502–509, Springer, Berlin, Heidelberg, 2005.

[39] L. Grady and G. Funka-Lea, "Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials," in *IEEE International Conference on Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis (ECCV 2004)*, pp. 230–245, Prague, Czech Republic: Springer, 2004.

[40] L. Grady, "Multilabel random walker image segmentation using prior models," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 763–770, San Diego, CA, USA, June 2005.

[41] L. Grady and A. K. Sinop, "Fast approximate random walker segmentation using eigenvector precomputation," in *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8, Anchorage, AK, USA, June 2008.

[42] S. Andrews, G. Hamarneh, and A. Saad, "Fast random walker with priors using precomputation for interactive medical image segmentation," in *IEEE International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2010)*, pp. 9–16, Springer, Berlin, Heidelberg, 2010.

[43] J. Goclawski, T. Weglinski, and A. Fabijanska, "Accelerating the 3D random walker image segmentation algorithm by image graph reduction and GPU computing," in *Conference on Image Processing and Communications Challenges 6. Advances in Intelligent Systems and Computing*, p. 313, Springer, Cham, 2015.

[44] C. H. Dong, Y. W. Chen, L. F. Lin et al., "Simultaneous segmentation of multiple organs using random walks," *Journal of Information Processing*, vol. 24, no. 2, pp. 320–329, 2016.

[45] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, vol. 2, pp. 23–26, Cambridge, UK, August 2004.

[46] "Visualization Toolkit," 2014, July 2011, http://www.vtk.org.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 9271251, 13 pages
https://doi.org/10.1155/2017/9271251

Research Article

Digital Path Approach Despeckle Filter for Ultrasound Imaging and Video

Marek Szczepański and Krystian Radlak

Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

Correspondence should be addressed to Marek Szczepański; marek.szczepanski@polsl.pl

Received 23 February 2017; Accepted 8 August 2017; Published 8 October 2017

Academic Editor: Junfeng Gao

We propose a novel filtering technique capable of reducing the multiplicative noise in ultrasound images that is an extension of the denoising algorithms based on the concept of digital paths. In this approach, the filter weights are calculated taking into account the similarity between pixel intensities that belongs to the local neighborhood of the processed pixel, which is called a path. The output of the filter is estimated as the weighted average of pixels connected by the paths. The way of creating paths is pivotal and determines the effectiveness and computational complexity of the proposed filtering design. Such procedure can be effective for different types of noise but fail in the presence of multiplicative noise. To increase the filtering efficiency for this type of disturbances, we introduce some improvements of the basic concept and new classes of similarity functions and finally extend our techniques to a spatiotemporal domain. The experimental results prove that the proposed algorithm provides the comparable results with the state-of-the-art techniques for multiplicative noise removal in ultrasound images and it can be applied for real-time image enhancement of video streams.

1. Introduction

Medical ultrasound is an imaging technique widely used in the diagnosis and assessment of internal body structures, and it plays a key role in treating various diseases. Compared to the other techniques of medical imaging, it is safe, noninvasive, and well tolerated by the patient, and ultrasound images are captured in real time at reasonable price. An accurate analysis of ultrasound images and thus an appropriate diagnosis are difficult due to the fact that the images are contaminated with characteristic granular structures called speckle noise, which deteriorates contrast and hinders the identification of important image details [1]. Although the ultrasound images are subjected to an initial improvement during the acquisition process, their quality is still far from optimal. Therefore, the main aim of image denoising is to remove the noise, while preserving the important details.

Recently, a plethora of methods capable of diminishing speckle noise have been proposed [2]. According to the works presented in [3–5], one of the most promising results for ultrasound images was obtained with algorithms based on anisotropic diffusion techniques [6] and the idea of nonlocal means [7–11]. However, majority of them were designed for static images, and not much attempt has been made to video by considering temporal coherence.

Video sequence processing algorithms can take an advantage of high correlation between adjacent frames, exploring spatial and temporal neighborhood. These properties are incorporated using different variants of averaging the pixel intensities in successive video frames. Simple temporal filters, such as temporal Gaussian, efficiently remove noise for videos with slowly moving objects, but averaging operation of subsequent frames can cause “ghosting” artifacts in the output results. Some works indicate that motion compensation allows us to reduce the blurring effect [12, 13]. Additionally, “ghosting” effects can be reduced by application of the statistical models that reflects the joint distributions of wavelet coefficients over space and time [14, 15].

Ghosting artifacts may be also omitted using temporal version of a bilateral filter that was successfully applied in the adaptive spatiotemporal accumulation (ASTA) filter [16].

In our previous works on video filtering, a spatiotemporal denoising scheme utilized local switching between spatial digital path approaches and temporal Gaussian filtering [17] and a fuzzy spatiotemporal filter that was later described [18]. Another interesting study connected with video denoising introduces a 3D filtering framework that is based on fuzzy set logic to combine the gradient values in different directions between previous and current temporal frames [19].

The main aim of this research is to develop a filter that will efficiently cope with multiplicative noise in ultrasound images and videos. The proposed algorithm is based on the idea of spatial digital paths presented in [20, 21]. The original 2D algorithm presented in [20] perfectly removes Gaussian noise and after some modifications impulsive noise but fails in the presence of multiplicative interferences. The new denoising scheme explores spatial or spatiotemporal pixel neighborhood in order to calculate the filter weights for pixels belonging to the processing window. Digital paths in a spatiotemporal domain could be understood as trajectories or object displacements in subsequent frames of a video stream.

The proposed technique utilizes a specific kind of digital paths, the so called *escaping paths*, and extends this concept from the spatial domain (2D) to the spatiotemporal domain (3D) that allows us to efficiently reduce the speckle noise. Furthermore, to increase the filter denoising ability, the new method of path creation utilizes the extended neighborhood introduced by von Neumann for cellular automata [22]. A detailed description of the proposed algorithm and its extensions are presented in Section 2. In Section 3, we present the results of experiments and the comparison with competitive filters. Finally, the conclusions are drawn in Section 4.

2. Method

2.1. Speckle Noise Model. Speckle noise is a signal-dependent and non-Gaussian multiplicative image distortion. Such noise is generally more difficult to remove than additive noise [23]. This type of distortion appears in sonar, laser, and synthetic aperture radar (SAR), and it depends on the structure of the material being imaged and various acquisition parameters [24]. Speckle noise is a random process and can be modeled using gamma, Rayleigh, and Fisher-Tippett distribution [25–27]. In our work, the simplified speckle noise model has been employed, since it has been applied successfully in many studies [11, 24, 28, 29].

$$u(\mathbf{x}) = \mathbf{v}(\mathbf{x}) + \sqrt{\mathbf{v}(\mathbf{x})} \cdot \mathbf{n}(\mathbf{x}), \quad (1)$$

where $u(\mathbf{x}), \mathbf{v}(\mathbf{x})$, and $\mathbf{n}(\mathbf{x})$ denote the observed signal, original unknown signal, and zero-mean Gaussian noise, respectively.

2.2. General Filter Framework. Since smoothing is commonly used to decrease the level of random noise, an averaging operation is required in order to replace the noisy pixel, $v(\mathbf{x})$, with a suitable pixel representative for the local spatiotemporal neighborhood of point $\mathbf{x} = (x, y, t)$. So the general form of the fuzzy adaptive filters in this work is defined as the weighted average of inputs inside the spatiotemporal window W that are in neighborhood relation $\mathcal{N}$ with a center pixel $\mathbf{x}$ [30, 31].

$$\widehat{v}(\mathbf{x}) = \sum_{\mathbf{x}_i \in \mathcal{N}(\mathbf{x})} w_i u(\mathbf{x}_i) = \frac{\sum_{\mathbf{x}_i \in \mathcal{N}(\mathbf{x})} \mu_i(\mathbf{x}, \mathbf{x}_i) u(\mathbf{x}_i)}{\sum_{\mathbf{x}_i \in \mathcal{N}(\mathbf{x})} \mu_i(\mathbf{x}, \mathbf{x}_i)}, \quad (2)$$

where $u(\mathbf{x}_i)$ and $\widehat{v}(\mathbf{x})$ denote filter inputs and outputs, respectively, and $\mu(\mathbf{x}, \mathbf{x}_i)$ is a similarity function computed along digital path starting at window central point $\mathbf{x}$, associated with its neighbor $\mathbf{x}_i$, and bounded by the spatiotemporal processing window W.

To describe the model of digital paths, a few notions should be introduced: digital lattice $\mathcal{H} = (X, \mathcal{N})$ defined by X, which is the set of all points of the image sequence, and a neighborhood relation $\mathcal{N}$ between the lattice points [32].

A digital path $P = \{p_i\}_{i=0}^{n}$ defined on the image lattice is a sequence of adjacent points $(p_i, p_{i+1}) \in \mathcal{N}$, and n is a number of path segments. Let $L(P)$ denote the length of the digital path $P\{p_i\}_{i=0}^{n}$ that is calculated as $\sum_{i=1}^{n-1} \rho(p_i, p_{i+1})$, where ρ denotes the spatiotemporal Euclidean distance between two adjacent points of the path.

The connection cost over the single digital path $P = \{p_i\}_{i=0}^{n}$ can be determined as a measure of dissimilarity between image pixels $p_0, p_1, \dots, p_n$ that forms a specific path linking p_0 and p_n [33, 34]. In the new approach, the connection cost will be calculated as a combination of the topological length of the path and the cumulative differences of gray levels. Thus, the connection cost for the entire path $\Lambda(P)$ can be defined as follows:

$$\Lambda(P) = \sum_{i=0}^{n-1} |u(p_{i+1}) - u(p_i)| \cdot \sum_{i=0}^{n-1} \rho(p_i, p_{i+1}). \quad (3)$$

2.3. Similarity Functions. Weights for the filter described by (2) can be defined in many ways; in our case, we use two different approaches based on connection costs calculated along digital paths. We created two kinds of membership functions which leads to the filters with slightly different properties.

2.3.1. DPA_{1st}. In this approach, similarity functions are defined for all neighbors of the central point $\mathbf{x}$ that remain in the neighborhood relation. To define the similarity function $\mu(\mathbf{x}, \mathbf{x}_i)$ between the filtered point $\mathbf{x}$ and its ith neighbor, we create digital paths starting at center point $\mathbf{x} = p_0$, intersecting $\mathbf{x}_i = p_1$ and finally terminating at point p_n, which may be reached in n steps from p_0. The similarity function defined for points p_0 and p_1 uses paths exploring the further neighborhood of the central point passing points $p_2, \dots, p_n$, so that the filtering result will be better suited to local image structures. An illustration of this idea is presented in Figure 1. This approach will

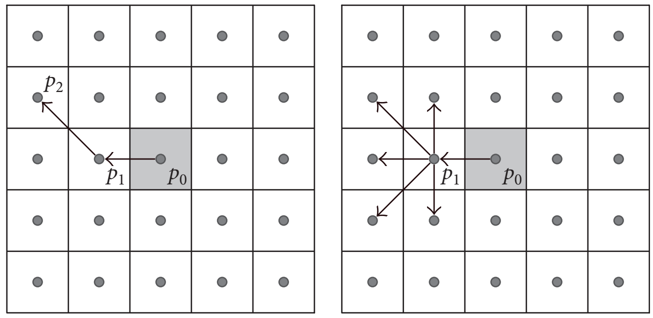


Figure 1: Illustration of paths created on the 2D image lattice with the $\text{DPA}_{1\text{st}}$ approach, used to determine the similarity function between two adjacent points.

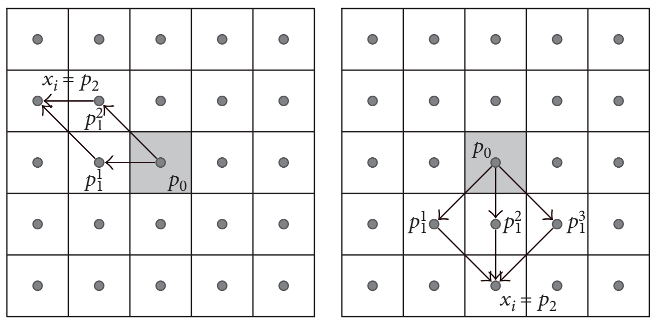


Figure 2: Illustration of paths created on the 2D image lattice with the DPA_{last} approach.

be further denoted as $\text{DPA}_{1\text{st}}$. In this case, the similarity function takes the form as follows:

$$\mu(\mathbf{x}, \mathbf{x}_i) = \mu(p_0, p_1) = \sum_{m=1}^{\omega} g(\Lambda\{p_0, p_1, p_2^m, \ldots, p_n^m\}), \quad (4)$$

where ω denotes the number of all possible paths $P_m = \{p_0, p_1, p_2^m, \ldots, p_n^m\}$ with n steps totally included in the processing window W, originating at $\mathbf{x} = p_0$ and crossing $\mathbf{x}_i = p_1$; m is the index of a specific path; $\Lambda\{\cdot\}$ is a dissimilarity value along a specific path; and $g(\cdot)$ is a smooth function of Λ. In this work, the exponential function is used as $g(\cdot)$ [20], so the similarity function takes the form as follows:

$$\mu(\mathbf{x}, \mathbf{x}_i) = \mu(p_0, p_1) = \sum_{m=1}^{\omega} \exp[-\beta \cdot \Lambda\{p_0, p_1, p_2^m, \ldots, p_n^m\}], \quad (5)$$

where β is the filter design parameter.

2.3.2. DPA_{last}. Another approach is to determine the similarity function between pixels $\mathbf{x}$ and $\mathbf{x}_i$ by all possible paths connecting them (Figure 2). This approach will be denoted as DPA_{last}, and the similarity function between image points $\mathbf{x}$ and $\mathbf{x}_i$ can be defined as follows:

$$\mu(\mathbf{x}, \mathbf{x}_i) = \mu(p_0, p_n) = \sum_{m=1}^{\omega} g(\Lambda\{p_0, p_1^m, p_2^m, \ldots, p_n\}), \quad (6)$$

where ω denotes the number of all possible paths $P_m = \{p_0, p_1^m, p_2^m, \ldots, p_n\}$ with n steps connecting $\mathbf{x}$ and $\mathbf{x}_i$ and totally included in the processing window W, $\Lambda\{\cdot\}$ is a dissimilarity value along a specific path, and $g(\cdot)$ is a smooth function of Λ. Finally, the DPA_{last} similarity function takes the form as follows:

$$\mu(\mathbf{x}, \mathbf{x}_i) = \mu(p_0, p_n) = \sum_{m=1}^{\omega} \exp[-\beta \cdot \Lambda\{p_0, p_1^m, p_2^m, \ldots, p_n\}]. \quad (7)$$

2.3.3. Filter Output and β Normalization. The proposed method can use different path lengths and bit depths, and therefore, to ensure the comparability of the results, it was necessary to rescale the parameter β. In this case, the β parameter will be divided by the maximum possible cost of the single path; thus, in (5), the normalized value $\widehat{\beta}$ will be used.

$$\widehat{\beta} = \frac{\beta}{\max(\Lambda(P_m))}. \quad (8)$$

Maximum path cost can be determined using the formula as follows:

$$\max(\Lambda(P_m)) = \left(2^{\text{bpp}} - 1\right) \cdot n^2 \cdot \sqrt{2}, \quad (9)$$

where bpp denotes the number of bits per pixel and n the number of path steps. The next stage is a normalization of the similarity function, which can be defined as follows:

$$\psi(\mathbf{x}, \mathbf{x}_i) = \frac{\mu(\mathbf{x}, \mathbf{x}_i)}{\sum_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x})} \mu\left(\mathbf{x}, \mathbf{x}_j\right)}. \quad (10)$$

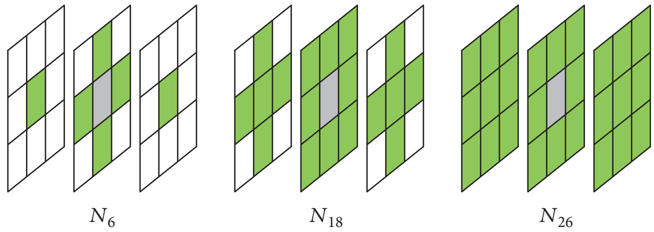


Figure 3: Basic spatiotemporal masks for different neighborhood systems.

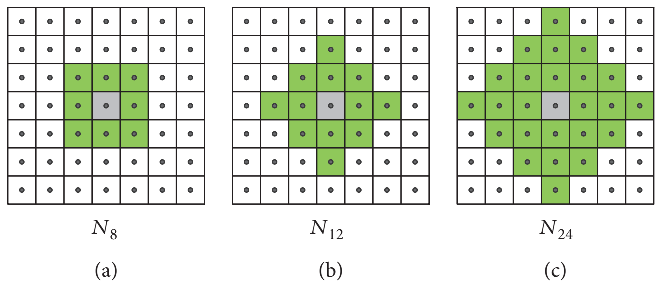


Figure 4: Spatial neighborhood systems utilized in our framework: (a) Standard 8 neighborhood. (b) and (c) show the extended von Neumann neighborhood with the radii 2 and 3.

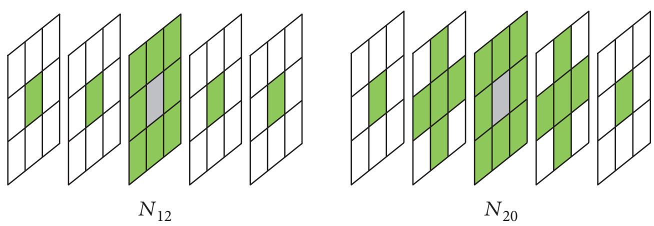


Figure 5: Modified spatiotemporal masks with extended neighborhood.

Let $\mathbf{x} = p_0$ denote the pixel under consideration, with $u(\mathbf{x}_i)$ representing the noisy pixel $\mathbf{x}_i$; the filter output $\hat{\mathbf{v}}(\mathbf{x})$ is given as follows:

$$\hat{v}(\mathbf{x}) = \sum_{\mathbf{x}_i \in \mathcal{N}(\mathbf{x})} \psi(\mathbf{x}, \mathbf{x}_i) \cdot u(\mathbf{x}_i). \tag{11}$$

2.4. Extended Neighborhood and Digital Path Models. The selected neighborhood system significantly affects the performance of the new filters. For the static images, there are two basic types of neighborhood: $\mathcal{N}_4$ and $\mathcal{N}_8$, while in the case of a three-dimensional image, three spatiotemporal neighborhood systems can be defined: $\mathcal{N}_6$, $\mathcal{N}_{18}$, and $\mathcal{N}_{26}$ (Figure 3). The new spatial filters with $\mathcal{N}_8$ neighborhood and the spatiotemporal ones with $\mathcal{N}_{18}$ and $\mathcal{N}_{26}$ are very efficient, but they fail for heavily degraded images. Therefore, in the proposed denoising design, we introduced the extended von Neumann neighborhood [22] originally defined for cellular automata. Various neighborhood systems for static 2D images are drawn in Figure 4, while 3D neighborhoods are shown in Figure 5.

Additionally, the efficiency of the proposed denoising framework is strongly connected with the type of digital paths. Different models of paths allow us to suppress a certain type of noise [20].

Previous research has demonstrated that the best results are obtained in the presence of impulsive Gaussian as well as of mixed-type noise for the *self-avoiding path* (SAP) model, but our experiments suggest that in the case of ultrasound images, the greatest results are achieved for the so called *escaping path model* (EPM); thus, the new filter will be denoted as the *escaping path filter* (EPF). In the proposed denoising scheme, the topological distance from the initial point in the following steps must be increased. Exemplary spatial escaping paths created with various neighborhood systems are illustrated in Figure 6, while Figure 7 depicts the spatiotemporal case. Later in this paper, the proposed filters will be marked as EPF2D for the case of two-dimensional filtering (2D) and EPF3D for the spatiotemporal case (3D).

In situations when the images are highly contaminated, we can increase the efficiency of the filter in one of two ways: (1) extend the length of the used paths or (2) apply it in an

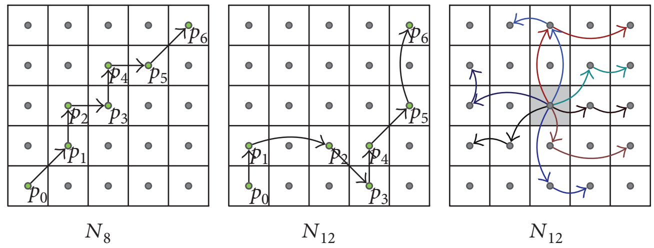


Figure 6: Escaping path model illustration with different neighborhood systems.

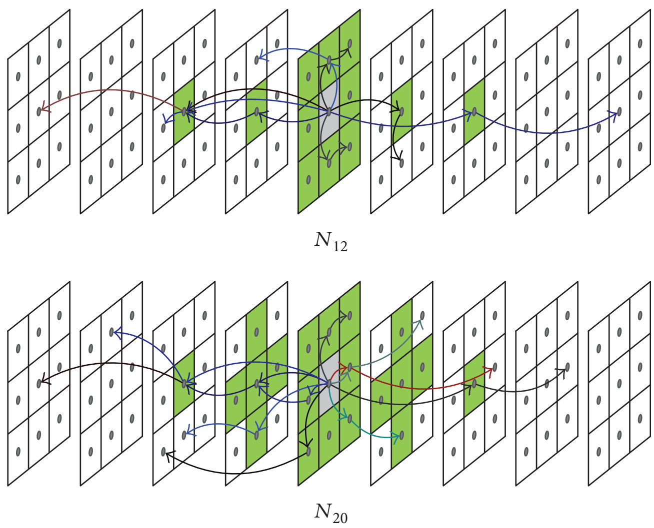


Figure 7: Examples of three-dimensional escaping path of length $n = 2$ limited by spatiotemporal window (spatial radius $A = 1$ and temporal radius $t = 4$).

iterative manner. The second option is much faster and more accurate and allows us to control the filter strength adjusting the β parameter in subsequent iterations. In this way, β can be updated as follows:

$$\beta(\kappa) = \beta(\kappa - 1) \cdot \alpha, \quad \kappa = 1, \ldots, m. \tag{12}$$

3. Simulation Results

3.1. Static Images. The commonly used benchmark images *goldhill*, *boats*, and artificially generated *phantom* were chosen to compare efficiency of different filters. Besides, we used a synthetic test image that was a phantom for a 3-month-old fetus denoted later as *fetus*. This test image was obtained using the *Field II* applications, which simulate the ultrasound field that is based on linear acoustics using the Tupholme-Stepanishen method for calculating the spatial impulse response [35–37]. The proposed ultrasound data seems to be more reliable, due to the fact that it contains artifacts typical for an ultrasound acquisition process and thus it can be used for image quality evaluation. A reference noise-free image was achieved by averaging of 500 simulated images and was depicted in Figure 8. The described filtering design has been compared with the following state-of-the-art methods capable of suppressing a speckle noise:

(i) Wiener filter [2]

(ii) Speckle reducing anisotropic diffusion (SRAD) [6]

(iii) Nonlocal means (NLM) [7, 8]

(iv) Optimized Bayesian nonlocal means (OBNLM) [11]

(v) Probabilistic nonlocal means (PNLM) [9]

(vi) Probabilistic patch-based weights (PPBW) [10]

(vii) Digital path approach ($\mathrm{DPA}_{\mathrm{last}}$) [20]

The source codes of the algorithms used in our comparison were provided by the authors of the respective papers. In order to determine the optimal values of parameters for the reference filters, we tested a wide range of parameters according to the authors' recommendations to obtain the highest possible PSNR value.

The recommended values of the parameters α and β for the proposed technique were adjusted, so that the best PSNR value was achieved. The test images were deteriorated by the multiplicative noise described by (1) with mean $\mu = 0$ and $\sigma^2 = 0.2, 0.4,$ and 0.6 except for the *fetus* image, which was

Goldhill Boats Phantom Fetus

Figure 8: Images used for the analysis.

Table 1: Comparison of the efficiency of the proposed filter and competitive methods using the PSNR and MSSIM quality measure on the static images.

File noise (σ^2)	Goldhill 0.2	Goldhill 0.4	Goldhill 0.6	Boats 0.2	Boats 0.4	Boats 0.6	Phantom 0.2	Phantom 0.4	Phantom 0.6	Fetus
PSNR results (dB)										
Noisy	31.52	19.69	13.25	30.65	18.75	12.62	35.05	23.03	16.48	22.46
Wiener	30.76	27.09	22.24	31.01	26.69	21.90	37.36	28.57	21.19	24.73
SRAD	30.73	27.26	23.69	32.06	26.96	23.98	42.52	30.82	24.14	18.62
NLM	34.06	27.35	16.49	35.40	27.68	15.54	43.07	30.31	20.15	27.27
OBNLM	30.87	28.24	24.43	32.61	28.06	24.13	38.64	30.34	23.76	18.78
PNLM	34.20	27.18	22.75	34.52	27.34	22.13	42.21	30.43	23.00	26.77
PPBF	32.80	27.11	18.56	32.64	26.91	20.00	37.01	29.50	21.15	25.93
DPA_{last}	32.83	27.77	24.96	32.61	27.10	24.20	44.98	34.65	26.63	26.69
EPF2D	33.63	27.98	25.24	33.48	27.49	24.21	48.57	37.60	27.32	27.34
MSSIM results										
Noisy	0.83	0.32	0.11	0.75	0.30	0.13	0.86	0.57	0.47	0.39
Wiener	0.77	0.64	0.43	0.84	0.67	0.43	0.98	0.91	0.67	0.66
SRAD	0.88	0.67	0.54	0.92	0.72	0.65	0.99	0.96	0.89	0.86
NLM	0.88	0.65	0.20	0.93	0.69	0.25	0.99	0.95	0.56	0.90
OBNLM	0.90	0.69	0.56	0.93	0.69	0.63	1.00	0.93	0.79	0.84
PNLM	0.90	0.64	0.42	0.92	0.67	0.39	0.98	0.91	0.80	0.76
PPBF	0.86	0.69	0.35	0.88	0.74	0.53	0.95	0.89	0.68	0.87
DPA_{last}	0.87	0.69	0.59	0.89	0.73	0.59	0.99	0.95	0.88	0.85
EPF2D	0.90	0.70	0.59	0.90	0.74	0.61	1.00	0.95	0.93	0.90

contaminated using a more realistic artifacts resulting from the physical model of an ultrasound image acquisition process. The restoration efficiency has been assessed using the peak signal-to-noise ratio (PSNR) and more sophisticated mean structural similarity index (MSSIM) calculated with Gaussian kernel and default parameters [38].

Numerical results obtained for the static images are summarized in Table 1. The conducted simulations revealed that the proposed EPF approach outperforms other techniques for highly deteriorated real images in terms of PSNR metric, while algorithms that are based on nonlocal means provide slightly better results for lower noise contamination level. Additionally, the proposed filter gives better results for synthetic images. A visual comparison of the results achieved for the phantom image is drawn in Figure 9. From this figure, it can be seen that most filters removed multiplicative noise and produces more visually pleasing results, but the outcomes are slightly blurred.

The most significant and valuable results seems to be obtained for simulated fetus image, because it is much similar to the realistic ultrasound images. The visual comparison of the performance of the analyzed filters for the *fetus* image is presented in Figure 10, while exemplary results for the real ultrasound images of fingers are depicted in Figure 11. The assessment of the achieved results can be also evaluated by segmentation accuracy applying image denoising as the preprocessing step, but currently, it is out of scope in this work. The proposed denoising scheme requires also a much

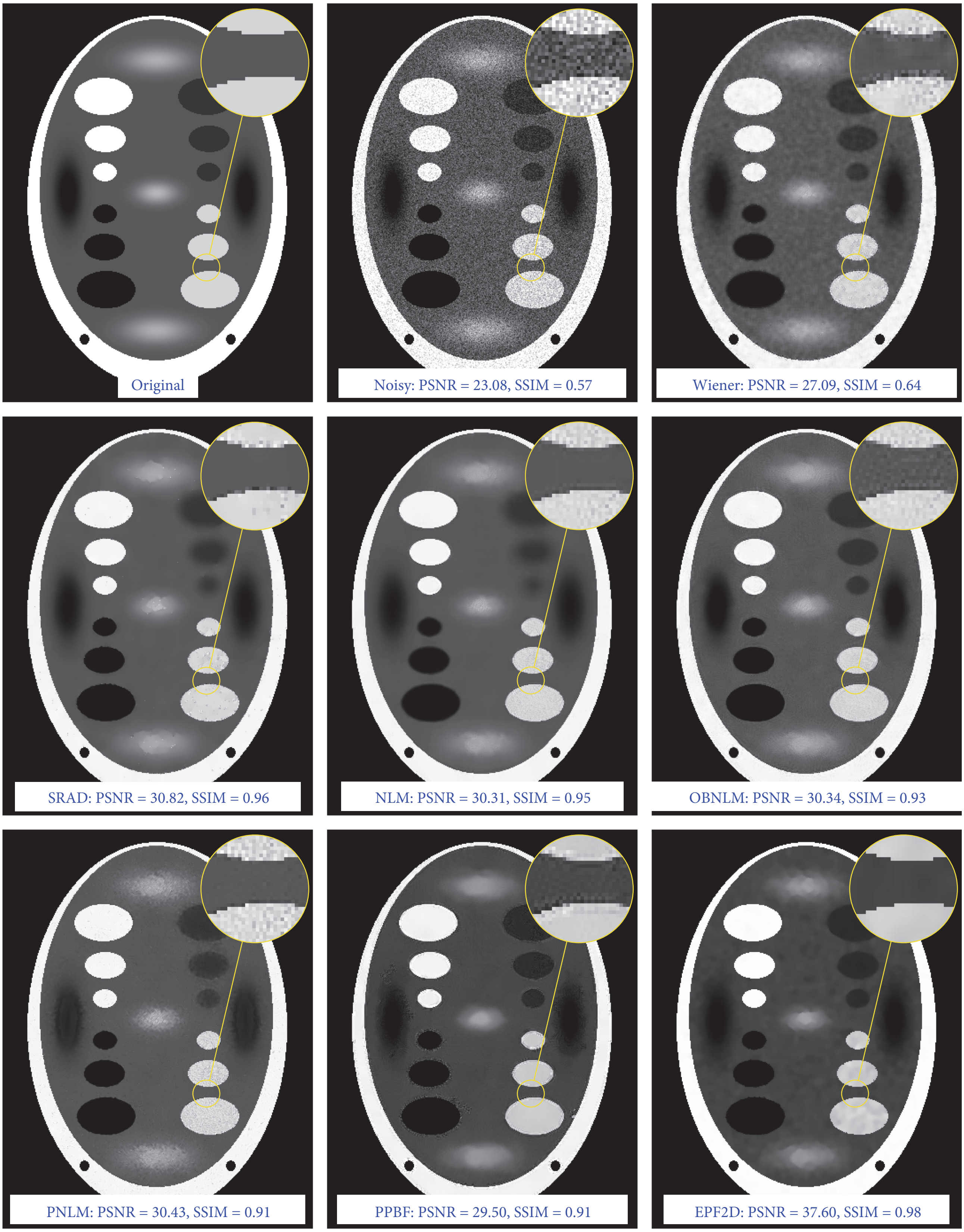


Figure 9: Visual comparison of the filtering efficiency of the artificial *phantom* image deteriorated by speckle noise with $\sigma^2 = 0.4$.

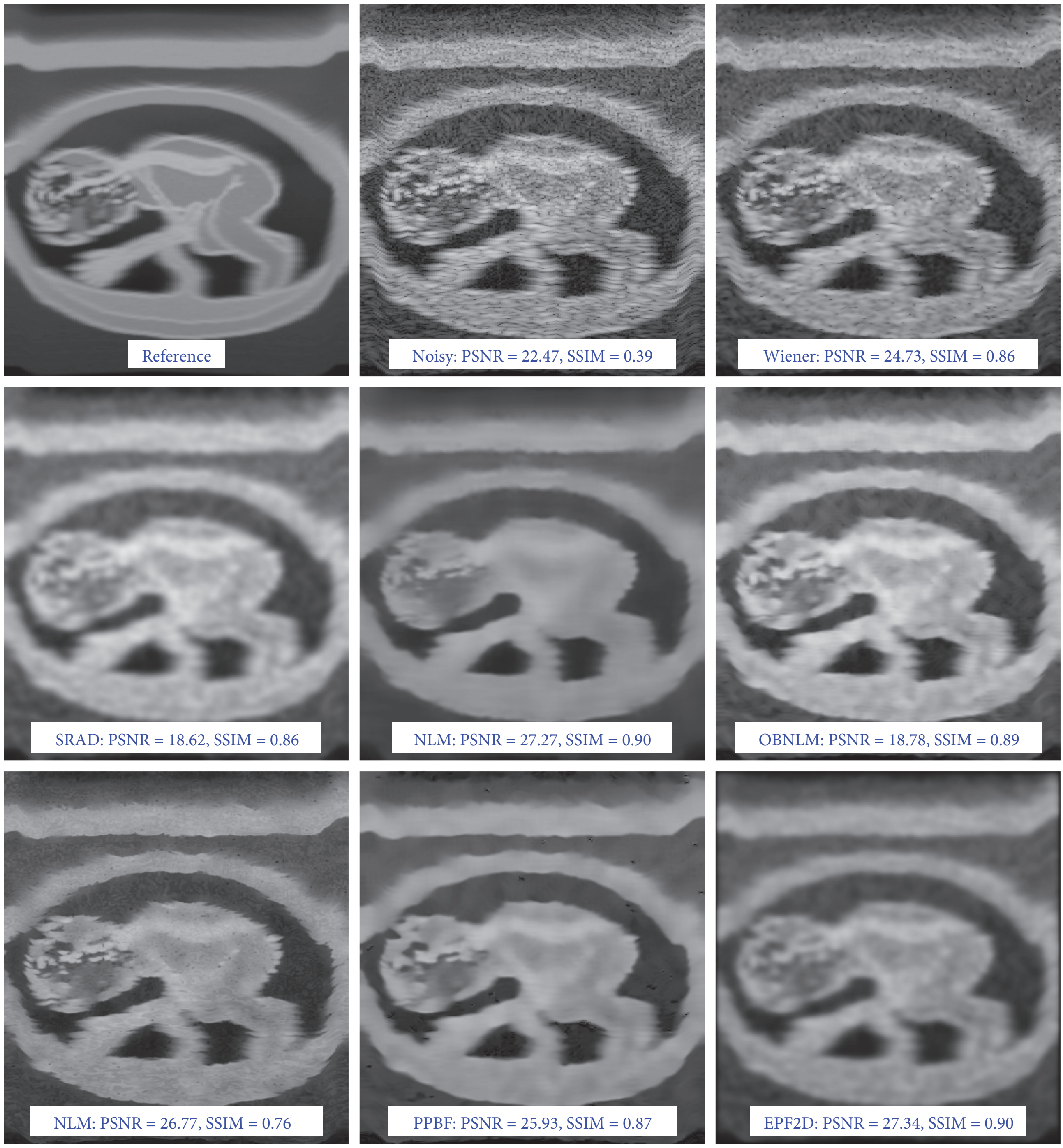


FIGURE 10: Visual comparison of the performance of the filtering algorithms for the *fetus* image obtained using the *Field II* application.

smaller neighborhood than that used in family of nonlocal means methods; therefore, our approach is much faster and less aggressively blurs the image. It should also be emphasized that for all methods based on the concept of NLM, even a small modification of optimal values of parameters gives a significant decrease in performance, while the EPF framework gives acceptable results for a wide spectrum of parameters.

3.2. Video Sequences. The images obtained by ultrasound devices are already enhanced by built-in filtering algorithms, but the resulting effect may still be unsatisfactory. So we can add another stage of image enhancement. Such technique should be capable of real-time processing. The *escaping path filters* are close to satisfy this condition on standard PC. Additionally, those filters are suitable for GPU implementation because all paths can be calculated in parallel and it lacks branches which block GPU threads. Our preliminary experiments show that we can obtain over 50 FPS for standard CIF sequences. Thus, several filters capable of real-time video processing were compared with our new

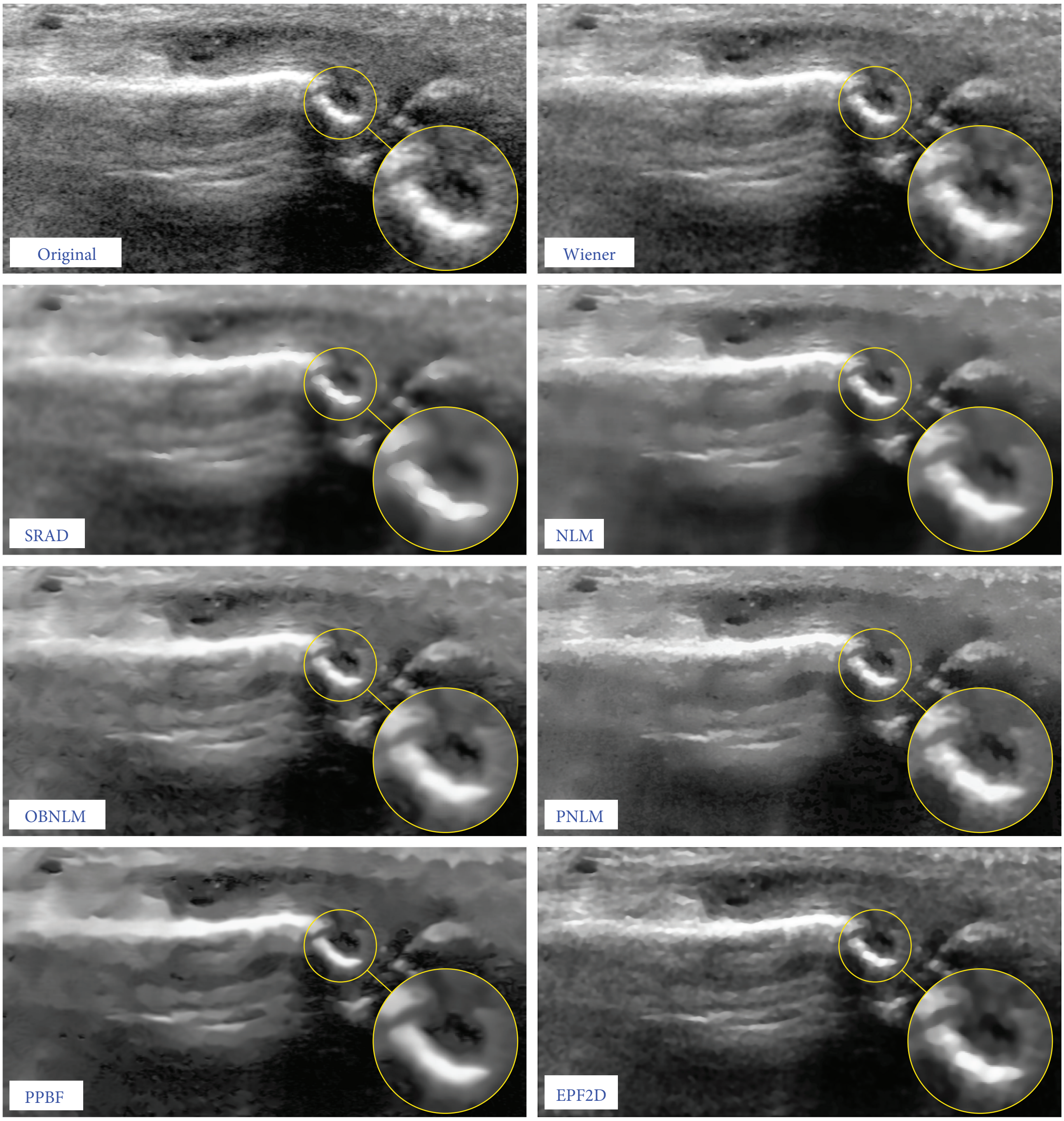


Figure 11: Visual comparison of the filtering efficiency evaluated on a real ultrasound image contaminated by multiplicative noise.

approach. Additionally more complex techniques, based on nonlocal means [8] and BM3D [39], were added to comparison; however, the computational complexity of those filters limits their use for offline processing. The performance of the following filters was evaluated:

(i) Wiener 2D—a spatially adaptive Wiener filter

(ii) Wiener 3D ($3 \times 3 \times 3$),

(iii) Temporal Gaussian filter (TGauss) ($n = 5$ and $\sigma = 5$)

(iv) Spatiotemporal median filter ($3 \times 3 \times 3$)

(v) Nonlocal means denoising [8] and our spatiotemporal implementation (NLM3D)

(vi) Block-matching and 3D filtering (BM3D) [39]

(vii) Spatial fast digital path approach (FDPA) ($\beta = 15$) [20]

(viii) Spatiotemporal fuzzy FDPA filter (STFFDPA) ($n = 5$, $\sigma = 5$, $\gamma = 4$, and $\beta = 15$) [18]

Table 2: Comparison of the filtering algorithms applied for standard test videos corrupted with different noise scenarios.

Video sequence noise σ^2	Foreman 0.2	Foreman 0.4	Foreman 0.6	Salesman 0.2	Salesman 0.4	Salesman 0.6	Tennis 0.2	Tennis 0.4	Tennis 0.6	Fetus
PSNR results (dB)										
Noisy	30.23	18.70	12.52	33.83	21.83	15.42	31.13	19.20	12.65	23.17
Wiener 2D	33.87	24.00	19.70	33.00	25.94	19.70	29.93	23.92	16.61	24.12
Wiener 3D	33.90	24.24	20.00	32.75	26.18	20.00	29.56	24.12	17.57	26.86
TGauss	30.60	23.30	20.61	33.56	26.50	20.61	27.56	22.50	17.32	26.30
Median 3D	30.00	26.65	24.51	29.15	27.48	24.51	21.70	21.21	19.64	26.36
NLM2D	35.05	27.92	22.28	34.37	26.88	22.28	25.96	26.39	17.75	29.10
NLM3D	35.67	28.02	22.34	32.79	27.06	22.34	30.75	27.64	17.81	30.28
BM3D	34.44	30.93	23.12	35.25	32.30	24.92	31.97	28.63	22.64	28.40
FDPA	32.74	25.32	20.23	31.28	27.24	20.24	27.40	24.09	16.73	24.80
STFFDPA	33.47	23.09	18.25	33.37	26.16	18.24	28.91	23.06	15.06	24.51
EPF2D	33.86	25.69	20.99	32.16	27.64	24.07	28.50	24.36	19.89	25.76
$\text{EPF}_{1\text{st}}$3D	34.39	28.05	25.02	34.40	29.28	25.02	30.85	27.20	20.99	28.92
EPF_{last}3D	33.27	28.10	25.05	34.28	29.07	25.06	30.11	25.85	21.20	29.34
MSSIM results										
Noisy	0.72	0.25	0.09	0.90	0.48	0.22	0.82	0.35	0.16	0.40
Wiener 2D	0.89	0.47	0.41	0.90	0.69	0.41	0.78	0.55	0.24	0.49
Wiener 3D	0.89	0.47	0.43	0.90	0.70	0.43	0.78	0.56	0.26	0.64
TGauss	0.81	0.40	0.41	0.92	0.67	0.41	0.82	0.49	0.25	0.61
Median 3D	0.84	0.63	0.59	0.86	0.76	0.59	0.59	0.47	0.35	0.75
NLM2D	0.92	0.71	0.50	0.91	0.73	0.50	0.85	0.68	0.30	0.90
NLM3D	0.92	0.71	0.50	0.88	0.73	0.50	0.86	0.69	0.31	0.91
BM3D	0.90	0.83	0.52	0.93	0.89	0.60	0.81	0.69	0.48	0.86
FDPA	0.87	0.55	0.42	0.88	0.74	0.42	0.72	0.54	0.24	0.54
STFFDPA	0.89	0.44	0.33	0.92	0.69	0.33	0.79	0.49	0.21	0.52
EPF2D	0.90	0.58	0.34	0.90	0.77	0.57	0.78	0.59	0.35	0.64
$\text{EPF}_{1\text{st}}$3D	0.91	0.72	0.64	0.93	0.83	0.63	0.86	0.69	0.40	0.88
EPF_{last}3D	0.89	0.71	0.60	0.93	0.82	0.60	0.84	0.65	0.41	0.89

(ix) Spatial escaping path filter $\text{EPM}_{1\text{st}}$2D ($\mathcal{N}_{12}$),

(x) Escaping path spatiotemporal filters $\text{EPM}_{1\text{st}}$3D and EPM_{last}3D ($\mathcal{N}_{20}$)

The video denoising algorithms were tested using publicly available video sequences: *foreman*, *salesman*, and *tennis*, contaminated by the multiplicative noise described in (1) with mean $\mu = 0$ and $\sigma^2 = 0.2, 0.4, \text{and}\, 0.6$. To obtain a more realistic comparison, we have prepared a test sequence based on the *fetus* image. Base fetus sequence consists of 150 frames subjected to different transformations that simulate the possible displacements during the ultrasound acquisition process. Then, all the frames have been subjected to a simulation using the *Field II* application. Virtually noise-free reference video was obtained by averaging of 200 simulation results (reference videos could be downloaded from http://dip.aei.polsl.pl/usg/fetus_video.7z). Table 2 summarizes results obtained for the set of test video sequences. Based on synthetic tests only, it is difficult to choose the best filter. In most cases, especially for small levels of noise, the best PSNR results are obtained for the BM3D filter; however, our spatiotemporal solution gives only slightly worse results at much higher processing speed. For heavily corrupted sequences, we can clearly see the advantage of the proposed filtrating technique. It should be noted that for the more realistic ultrasound noise model, obtained using the *Field II* application, the advantages of our solution is clear (the best results were obtained for the NLM3D filter, but the computational complexity disqualifies it entirely, even for offline processing). In the case of the *tennis* sequence with minimal disruption, most filtering techniques deteriorate quality ratios. This is due to the specific background that resembles an impulsive noise pattern. Figure 12 presents exemplary filtering results and SSIM maps for the mentioned sequence. It is thus clear that the worst results were obtained for filtering based on the median, which effectively removes elements of background texture. In addition, median 3D filtering introduces some temporal artifacts, such as blurred or erased moving objects. It should be noted that the ranking obtained with the SSIM index is slightly different in favor of the escaping path filters.

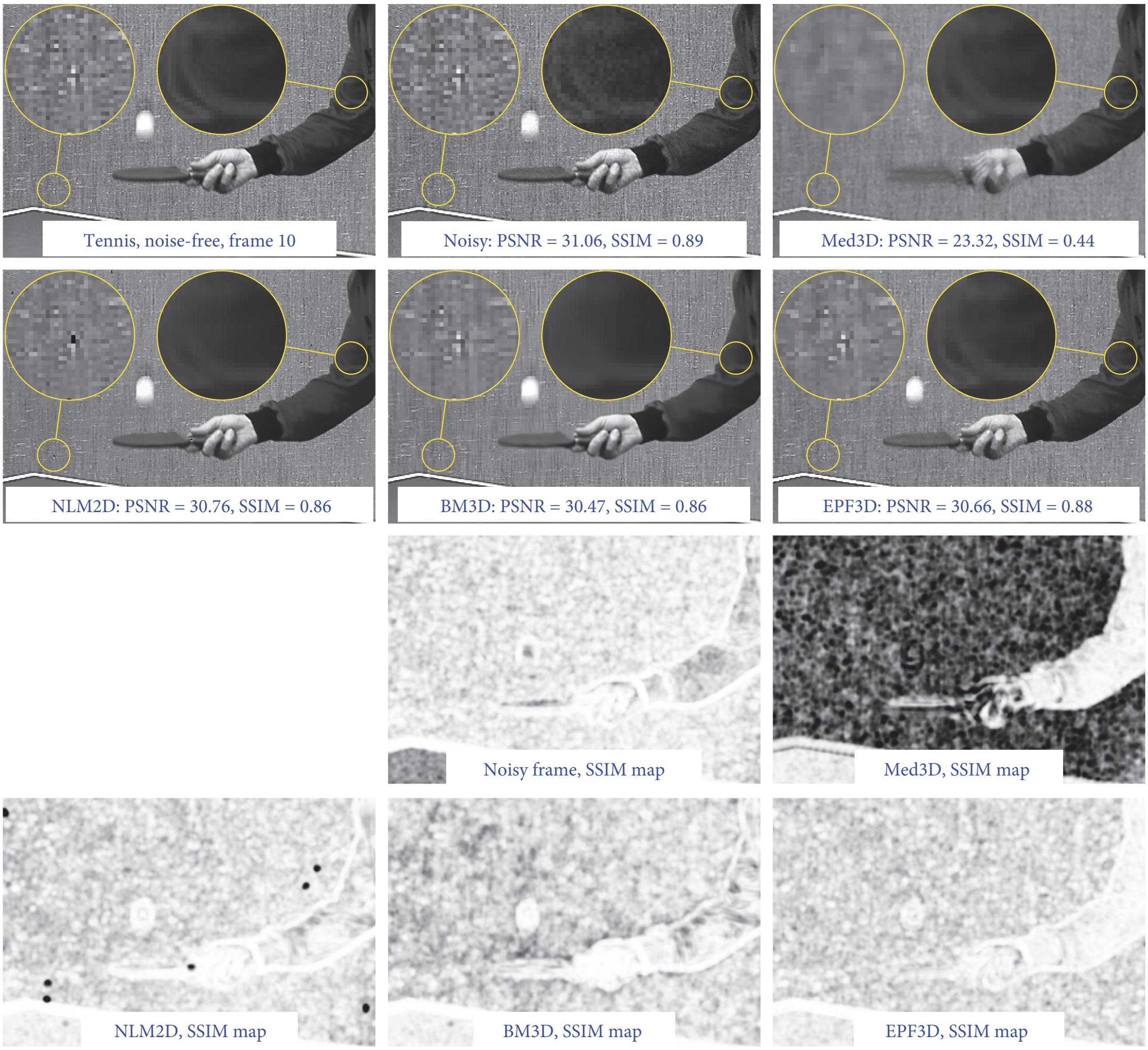


FIGURE 12: Denoising results of frame 10 from the *tennis* video sequence corrupted with multiplicative noise ($\sigma^2 = 0.2$) and the corresponding SSIM quality maps.

4. Conclusions

In this paper, a new class of fast spatial and spatiotemporal filters was presented. The proposed filtering techniques were designed for multiplicative noise suppression, specifically for ultrasound image and video filtering. The novel approach is based on the special type of digital paths, the so called *escaping paths*, created on an image lattice—spatial in a 2D case or spatiotemporal for video processing. Additionally, the new extended neighborhood model was introduced, based on von Neumann concept derived from cellular automata theory. The presented methods give comparable or better results to the other methods, both for static image and video sequences. Another beneficial feature of the proposed denoising scheme is its lower computational complexity than that of other state-of-the-art techniques, which allows us to apply it in real-time image processing tasks. The proposed filter is also more stable for a wide range of input parameters and gives satisfactory results in terms of different quality metrics and visual inspection.

Conflicts of Interest

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

Acknowledgments

This project was partially financially supported by the SUT Grant BK/213/Rau1/2016/10.

References

[1] F. Latifoğlu, "A novel approach to speckle noise filtering based on artificial bee colony algorithm: an ultrasound image application," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 3, pp. 561–569, 2013.

[2] C. P. Loizou, C. Theofanous, M. Pantziaris, and T. Kasparis, "Despeckle filtering software toolbox for ultrasound imaging of the common carotid artery," *Computer Methods and Programs in Biomedicine*, vol. 114, no. 1, pp. 109–124, 2014.

[3] K. Radlak and B. Smolka, "Adaptive non-local means filtering for speckle noise reduction," in *Computer Vision and Graphics*, L. Chmielewski, R. Kozera, B. S. Shin and K. Wojciechowski, Eds., vol. 8671 of Lecture Notes in Computer Science, Springer International Publishing, Cham, 2014.

[4] M. Szczepanski, K. Radlak, and A. Popowicz, *Escaping Path Approach with Extended Neighborhood for Speckle Noise Reduction*, Springer International Publishing, Cham, 2015.

[5] A. F. de Araujo, C. E. Constantinou, and J. M. R. Tavares, "Smoothing of ultrasound images using a new selective average filter," *Expert Systems with Applications*, vol. 60, pp. 96–106, 2016, http://www.sciencedirect.com/science/article/pii/S0957417416302081.

[6] Y. Yu and S. T. Acton, "Speckle reducing anisotropic diffusion," *IEEE Transactions on Image Processing*, vol. 11, no. 11, pp. 1260–1270, 2002.

[7] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 60–65, San Diego, CA, USA, 2005.

[8] A. Buades, B. Coll, and J. M. Morel, "Non-local means denoising," *Image Processing On Line*, vol. 1, 2011.

[9] Y. Wu, B. Tracey, P. Natarajan, and J. P. Noonan, "Probabilistic non-local means," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 763–766, 2013.

[10] C. A. Deledalle, L. Denis, and F. Tupin, "Iterative weighted maximum likelihood denoising with probabilistic patch-based weights," *IEEE Transactions on Image Processing*, vol. 18, no. 12, pp. 2661–2672, 2009.

[11] P. Coupe, P. Hellier, C. Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," *IEEE Transactions on Image Processing*, vol. 18, no. 10, pp. 2221–2229, 2009.

[12] E. Dubois and S. Sabri, "Noise reduction in image sequences using motion-compensated temporal filtering," *IEEE Transactions on Communications*, vol. 32, no. 7, pp. 826–831, 1984.

[13] G. Varghese and Z. Wang, "Video denoising based on a spatiotemporal Gaussian scale mixture model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 7, pp. 1032–1040, 2010.

[14] N. X. Lian, V. Zagorodnov, and Y. P. Tan, "Video denoising using vector estimation of wavelet coefficients," in *2006 IEEE International Symposium on Circuits and Systems*, pp. 2673–2676, Island of Kos, Greece, 2006.

[15] S. M. M. Rahman, M. O. Ahmad, and M. N. S. Swamy, "Video denoising based on inter-frame statistical modeling of wavelet coefficients," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 2, pp. 187–198, 2007.

[16] E. P. Bennett and L. McMillan, "Video enhancement using per-pixel virtual exposures," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 845–852, 2005.

[17] M. Szczepanski, "Spatio-temporal filters in video stream processing," in *Computer Recognition Systems 4*, R. Burduk, M. Kurzyński, M. Woźniak and A. Żołnierek, Eds., vol. 95 of Advances in Intelligent and Soft Computing, Springer, Berlin Heidelberg, 2011.

[18] M. Szczepanski, "Spatio-temporal fuzzy FDPA filter," in *Computer Analysis of Images and Patterns*, P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano and W. Kropatsch, Eds., vol. 6855 of Lecture Notes in Computer Science, pp. 316–323, Springer, Berlin Heidelberg, 2011.

[19] V. Ponomaryov, H. Montenegro, A. Rosales, and G. Duchen, "Fuzzy 3D filter for color video sequences contaminated by impulsive noise," *Journal of Real-Time Image Processing*, vol. 10, no. 2, pp. 313–328, 2012.

[20] M. Szczepanski, B. Smolka, K. Plataniotis, and A. Venetsanopoulos, "On the geodesic paths approach to color image filtering," *Signal Processing*, vol. 83, no. 6, pp. 1309–1342, 2003.

[21] M. Szczepanski, B. Smolka, K. Plataniotis, and A. Venetsanopoulos, "On the distance function approach to color image enhancement," *Discrete Applied Mathematics*, vol. 139, no. 1–3, pp. 283–305, 2004.

[22] J. V. Neumann, *Theory of Self-Reproducing Automata*, University of Illinois Press, Champaign, IL, USA, 1966.

[23] A. Achim, A. Bezerianos, and P. Tsakalides, "Ultrasound image denoising via maximum a posteriori estimation of wavelet coefficients," in *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2001*, vol. 3, pp. 2553–2556, Istanbul, Turkey, 2001.

[24] M. Hacini, F. Hachouf, and K. Djemal, "A new speckle filtering method for ultrasound images based on a weighted multiplicative total variation," *Signal Processing*, vol. 103, pp. 214–229, 2014, Image Restoration and Enhancement: Recent Advances and Applications.

[25] R. F. Wagner, S. W. Smith, J. M. Sandrik, and H. Lopez, "Statistics of speckle in ultrasound B-scans," *IEEE Transactions on Sonics and Ultrasonics*, vol. 30, no. 3, pp. 156–163, 1983.

[26] Z. Tao, H. D. Tagare, and J. D. Beaty, "Evaluation of four probability distribution models for speckle in clinical cardiac ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 25, no. 11, pp. 1483–1491, 2006.

[27] G. Slabaugh, G. Unal, T. Fang, and M. Wels, "Ultrasound-specific segmentation via decorrelation and statistical region-based active contours," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 45–53, New York, NY, USA, 2006.

[28] T. Loupas, W. N. McDicken, and P. L. Allan, "An adaptive weighted median filter for speckle suppression in medical ultrasonic images," *IEEE Transactions on Circuits and Systems*, vol. 36, no. 1, pp. 129–135, 1989.

[29] K. Krissian, R. Kikinis, C. F. Westin, and K. Vosburgh, "Speckle-constrained filtering of ultrasound images," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 547–552, San Diego, CA, USA, 2005.

[30] K. Plataniotis, D. Androutsos, and A. Venetsanopoulos, "Fuzzy adaptive filters for multichannel image processing," *Signal Processing*, vol. 55, no. 1, pp. 93–106, 1996.

[31] K. N. Plataniotis, D. Androutsos, and A. N. Venetsanopoulos, "Adaptive fuzzy systems for multichannel signal processing," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1601–1622, 1999.

[32] M. Schmitt, "Lecture notes on geodesy and morphological measurements," in *Proceedings of the Summer School on Morphological Image and Signal Processing*, pp. 36–91, Zakopane, Poland, 1995.

[33] O. Cuisenaire, *Distance Transformations: Fast Algorithms and Applications to Medical Image Processing. [Ph.D. thesis]*, Universite Catholique de Louvain, Belgium, 1999.

[34] P. J. Toivanen, "New geodesic distance transforms for gray-scale images," *Pattern Recognition Letters*, vol. 17, no. 5, pp. 437–450, 1996.

[35] J. Jensen and N. Svendsen, "Calculation of pressure fields from arbitrarily shaped, apodized, and excited ultrasound transducers," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 39, no. 2, pp. 262–267, 1992.

[36] J. A. Jensen, "Field: a program for simulating ultrasound systems," *10th Nordic-Baltic Conference on Biomedical Imaging*, vol. 4, Supplement 1, Part 1, pp. 351–353, 1996.

[37] J. A. Jensen and P. Munk, *Computer Phantoms for Simulating Ultrasound B-Mode and CFM Images*, Springer US, Boston, MA, 1997.

[38] 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.

[39] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 8314740, 7 pages
https://doi.org/10.1155/2017/8314740

Research Article

Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images

QingZeng Song, Lei Zhao, XingKe Luo, and XueChen Dou

School of Computer Science & Software Engineering, Tianjin Polytechnics University, Tianjin, China

Correspondence should be addressed to Lei Zhao; 1976958227@qq.com

Received 10 March 2017; Accepted 14 May 2017; Published 9 August 2017

Academic Editor: Junfeng Gao

Lung cancer is the most common cancer that cannot be ignored and cause death with late health care. Currently, CT can be used to help doctors detect the lung cancer in the early stages. In many cases, the diagnosis of identifying the lung cancer depends on the experience of doctors, which may ignore some patients and cause some problems. Deep learning has been proved as a popular and powerful method in many medical imaging diagnosis areas. In this paper, three types of deep neural networks (e.g., CNN, DNN, and SAE) are designed for lung cancer calcification. Those networks are applied to the CT image classification task with some modification for the benign and malignant lung nodules. Those networks were evaluated on the LIDC-IDRI database. The experimental results show that the CNN network archived the best performance with an accuracy of 84.15%, sensitivity of 83.96%, and specificity of 84.32%, which has the best result among the three networks.

1. Introduction

Lung cancer which is the most common cancer in both men and women is a major burden of disease worldwide [1]. Some report estimated that the number of new cases of lung cancer is about 221,200, accounting for about 13% of all cancer diagnoses in 2015. The mortality of lung cancer accounts for about 27% of all cancer deaths [2]. For those reasons, lung nodules need to be examined and watched closely when it might be at an early stage. By the early detection, the 5-year survival rate of patients with lung cancer can be improved by about 50%.

Computed tomography (CT) is the most effective method of lung nodule detection for its ability to form three-dimensional (3D) images of the chest, resulting in greater resolution of nodules and tumor pathology. A CT image by computer processing to assist lung nodule diagnostics has been widely used in clinic. The process of computer-aided diagnosis (CAD) of lung cancer can be divided into a detection system (often abbreviated as CADe) and diagnostic system (often abbreviated as CADx). The CADe system divides the candidate nodules identified in the previous step into nodules or nonnodules (i.e., normal anatomic structures). The goal of the CADx system is to classify detected nodules into benign and malignant nodules [3]. Since the probability of malignancy is closely related to the geometric size, shape, and appearance, CADx can distinguish the benign and malignant pulmonary nodules by the effective features such as texture, shape, and growth rate. Thus, the success of a particular CADx system can be measured in terms of accuracy of diagnosis, speed, and automation level [4].

In recent years, neural networks, rebranded as "deep learning," began beating traditional AI in every critical task: recognizing speech; characterizing images; and generating natural, readable sentences. Deep learning not only accelerates the critical task but also improves the precision of the computer and the performance of CT image detection and classification.

In this paper, the problem of classification of benign and malignant is considered. It is proposed to employ, respectively, the convolution neural network (CNN), deep neural network (DNN), and stacked autoencoder (SAE). The work can be used as input directly to reduce the complex

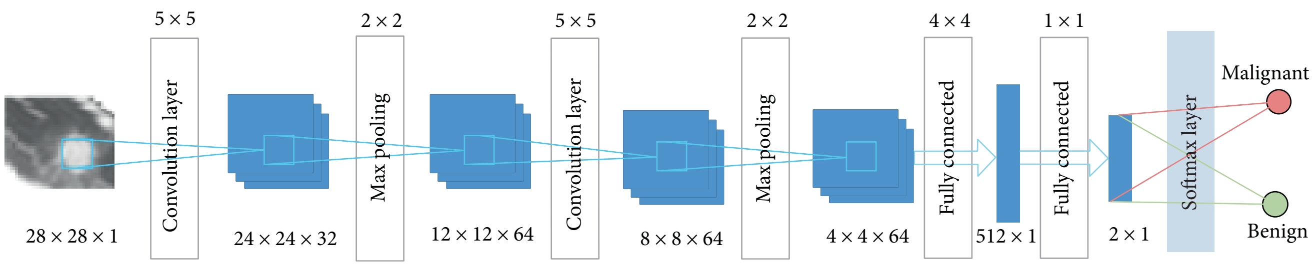


Figure 1: The architecture of the CNN.

reconstruction of data in the process of feature extraction and classification.

The rest of the paper is organized as follows. Section 2 analyzes the related works. Section 3 presents the proposed methodology for the classification of lung nodules. The experimental results obtained are discussed in Section 4. The conclusion of this paper was made in Section 5.

2. Related Works

Various initiatives are frequently developed aiming at increasing the accuracy of lung cancer diagnosis using a neural network. Chen et al. [5] proposed a method that uses a neural network ensemble (NNE) scheme to distinguish probably benign and uncertain and probably malignant lung nodules. Experimental results illustrated that the scheme had classification accuracy (78.7%) which is better than that of the individual classifier (LVQNN: 68.1%).

In [6], Kuruvilla and Gunavathi proposed a methodology based on texture features using the artificial neural network (ANN), with an accuracy rate of 93.30%. Using the combination of texture and shape features for detection and classification may result in improved classification accuracy [7]. Kumar et al. presented a methodology using the stacked autoencoder (SAE), a deep learning technique, with an accuracy rate of 75.01% [8].

Deep learning is based on using "deep" neural networks comprised of a large number of hidden layers. The deep belief network (DBN) which has undirected connections between its top two layers and downward-directed connections between all its lower layers [9] has been tested for classification of malignancy of lung nodules without computing the morphology and texture features [10]. It had reached the sensitivity rate of 73.40% and the specificity rate of 82.20% using the deep belief network.

Some research papers applied deep CNNs for detection or classifications of a medical image. In 2015, Shen et al. [11] diagnosed lung cancer on the LIDC database using a multiscale two-layer CNN and the reported accuracy was 86.84%. In [12], Shin et al. exploit and extensively evaluate three important, previously understudied factors on CNN architecture, dataset characteristics, and transfer learning.

3. Materials and Methods

In this section, the proposed approach on the LIDC-IDRI [13] dataset from the Lung Image Database Consortium is evaluated. The complex steps of image feature extraction in traditional medicine can be reduced by directly inputting the original image.

Table 1: Parameter of the CNN.

Layer	Type	Input	Kernel	Output
1	Convolution	28 × 28 × 1	5 × 5	24 × 24 × 32
2	Max pooling	24 × 24 × 32	2 × 2	12 × 12 × 64
3	Convolution	12 × 12 × 64	5 × 5	8 × 8 × 64
4	Max pooling	8 × 8 × 64	2 × 2	4 × 4 × 64
5	Fully connected	4 × 4 × 64	4 × 4	512 × 1
6	Fully connected	512 × 1	1 × 1	2 × 1
7	Softmax	2 × 1	N/A	Result

3.1. Convolution Neural Networks (CNNs). A convolution neural network (CNN) is a multilayer neural network, which comprised of one or more convolution layers and then followed by one or more fully connected layers as in a standard multilayer neural network. The CNN was proposed in 1960s, with the ideas like local perception, the weights of sharing, and sampling in space or time. Local perception can find some local characteristics of the data for the basic features of the visual animals, such as an angle and an arc in the picture [14]. It is a kind of an efficient identification method which has attracted wide attention recently. The benefit of CNNs is that they are easier to train and have many fewer parameters than fully connected networks with the same number of hidden units.

Convolution neural network architecture is usually used in collaboration with the convolution layer and pool layer [15]. The affection of the pooling layer is to confuse the features of the specific position. Since some location features are not important, it just needs other features and the relative position. The pooling layer operation consists of max pooling and mean pooling. Mean pooling calculates the average neighborhood within the feature points, and max pooling calculates the neighborhood within a maximum of feature points. The error of feature extraction mainly comes from two aspects: the neighborhood size limitation caused by the estimated variance and convolution layer parameter estimated error caused by the mean deviation. Mean pooling can reduce the first error, retaining more image background information. Max pooling can reduce the second error, retaining more texture information.

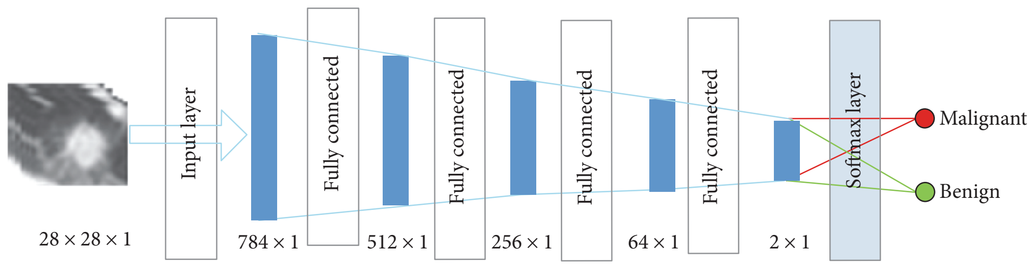


Figure 2: The architecture of the DNN.

Table 2: Parameter of the DNN.

Layer	Type	Input	Output
1	Input	$28 \times 28 \times 1$	784×1
2	Fully connected	784×1	512×1
3	Fully connected	512×1	256×1
4	Fully connected	256×1	64×1
5	Fully connected	64×1	2×1
6	Softmax	2×1	Result

The architecture of the CNN in this paper is showed in Figure 1. It is composed of multiple maps in each layer; each map is composed of multiple neural units, all the neural units in the same map share one convolution kernel (i.e., weight), and each convolution kernel represents a feature, such as access to the edge of image features. The detail of the CNN is showed in Table 1. The input data (image data) has a strong robustness on the distortion. The multiscale convolution image feature is generated by setting the convolution kernel size and parameter; the information of different angles is generated in the feature space.

3.2. Deep Neural Network (DNN). A DNN is an increase in the number of hidden nodes in a simple neural network. The neural network can be used to carry on the more complex input calculation, because each hidden layer can be the nonlinear transformation of the output layer and the deep neural network is better than the "shallow" network. The nonlinear $f(x)$ should be used for each hidden layer, because if the activation function is linear, compared with the single hidden layer neural network, the depth of the hidden layer of the network does not enhance the ability to express. The processing part of the pulmonary nodule is decomposed into the DNN, so that different network layers can be used to obtain the characteristics of the pulmonary nodules with different sizes. There are also local extremum problems and gradient diffusion problems in the DNN.

In the training process, the original image is used as the input layer parameters, so as to retain a large amount of detailed information of the image. The input layer, hidden layer, and output layer of the DNN architecture are all connected layers, and the DNN does not contain a convolution layer. DNN training images and label was input into the DNN architecture; each layer of the weight in the first training is randomly

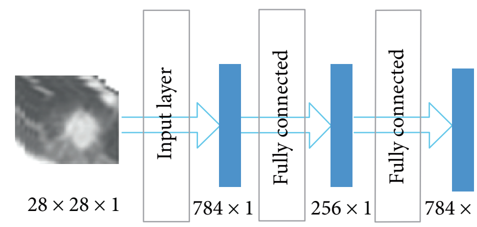


Figure 3: Sparse autoencoder.

generated by Gauss distribution, setting the bias to 0. Then, the output value calculated is the forward propagation and update parameters are the back propagation. The depth of the neural network structure is in Figure 2 and is further detailed in Table 2. Because the parameters of DNN are too prone to overfitting [16], fine-tuning [17], increasing the data volume, and regularization [18] are needed to solve it.

3.3. Stacked Autoencoder (SAE). A stacked autoencoder (SAE) neural network is a multilayer sparse autoencoder of a neural network. The sparse autoencoder is an unsupervised learning algorithm [19]. The sparse autoencoder is divided into three layers, namely, the input layer, hidden layer, and output layer. The number of neurons in the input and output layers is the same, and the number of hidden neurons is less than that of the input layer. Figure 3 is the structure of the sparse autoencoder. In addition, the sparse autoencoder is divided into a coding stage and decoding stage; the coding stage is the mapping of the input layer to the hidden layer. The decoding phase is the mapping of the hidden layer to the output layer. In this paper, multiple autoencoders and softmax classifiers are combined to construct a SAE network with multiple hidden layers and a final softmax classifier [20].

Figure 4 is the structure of the stacked autoencoder neural network. The hidden layer is the hidden layer of a single sparse autoencoder. The diagnosis of lung nodules belongs to the problem of image classification; each sparse autoencoder deletes the "decode" layer after the training is completed and directly uses the encoding process for the next sparse autoencoder training of the output.

3.4. Loss Functions of the Neural Network. The loss function is as follows:

$$C(w,b) \equiv \frac{1}{2n}\sum_{x}\|y(x)-a\|^{2} + \frac{1}{2n}\lambda\sum_{w}w^{2}, \qquad (1)$$

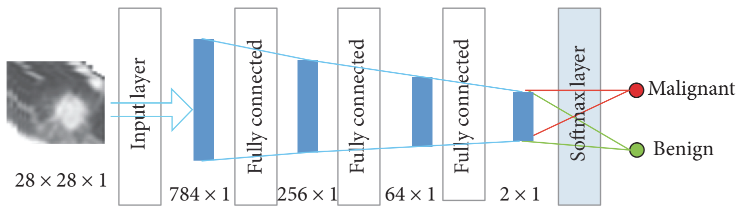
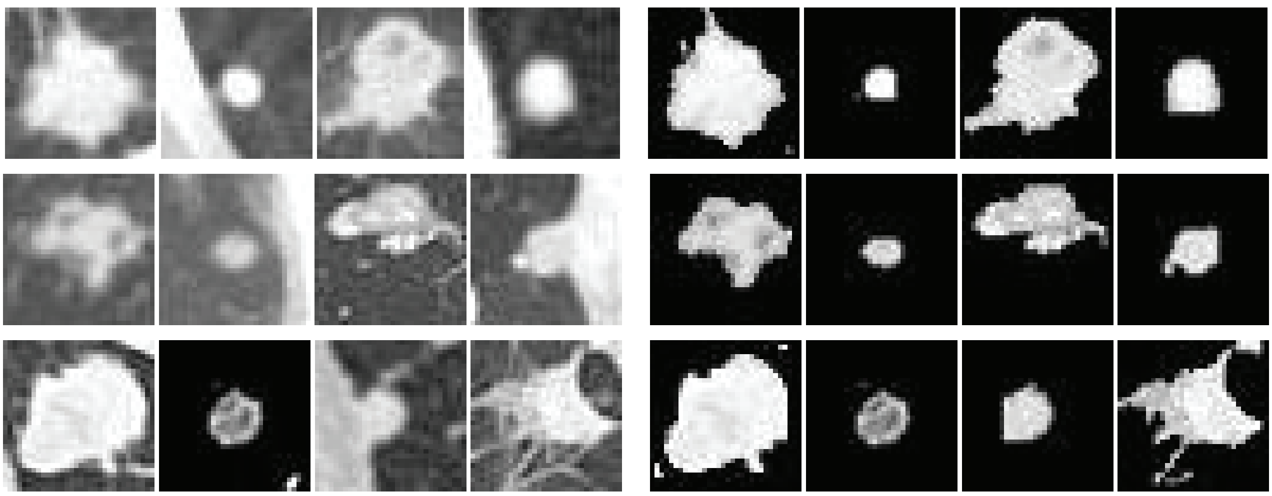


Figure 4: Architecture of the SAE.



(a) (b)

Figure 5: Nodular images.

where C is the cost function, w is the weight, b is the bias, n is the number of training dataset instances, x is the image pixel values as an input parameter, and a is the output value. The DNN is used to carry on the back propagation operation to modify the weight w and paranoid b, so that the difference between the predicted value and the real value is getting smaller and smaller, and thus, the accuracy is improved. The last item of the loss function is to prevent overfitting in the training process, and the sum of all weights is divided by $2n$. Another method to prevent overfitting is dropout, which randomly shields some neurons before the back propagation, and the masked neurons do not update the parameters. Since the DNN needs a lot of data, but if a large number of data are input into the neural network, it requires a lot of memory. Therefore, in order to modify the parameters more quickly, every time a min_batch to do a back propagation.

The activation function of the neural network is Leaky ReLU, which can enhance the ability of nonlinear modeling. The ReLU activation function formula is as follows:

$$y = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases} \tag{2}$$

where x is the result of weighted priority multiplication and paranoid addition and y is the output of the activation function. It can be seen that the derivative of ReLU is 0 if $x < 0$, else 1. So ReLU eliminates the problem of the gradient of the sigmoid activation function. However, with the continuous updating of the training, the weight cannot continue to be updated, which is known as "the phenomenon of neuronal death." On the other hand, the output of ReLU is more than 0, that is, the output of the neural network is offset. The above problems can be solved using Leaky ReLU. The Leaky ReLU activation function formula is as follows:

$$y = \begin{cases} x & \text{if } x \geq 0 \\ ax & \text{if } x < 0, \end{cases} \tag{3}$$

where a is set to 0.1; a in Leaky ReLU is fixed and in the ReLU is not fixed.

3.5. LIDI-IDRI. The database used in this paper is LIDC-IDRI, which contains 244,527 images of the 1010 cases. Each subject includes images from a clinical thoracic CT scan and an associated XML file that records the results of a two-phase image annotation process performed by four experienced thoracic radiologists [13]. The distribution of thickness of CT images in lung nodules is extensive. Most of them are concentrated at 1 mm, 1.25 mm, and 2.5 mm. The size of the patient's pulmonary nodules is from 3 mm to 30 mm. The number of benign nodules with small diameter is larger, and the number of malignant nodules with larger diameter is smaller. But it is not sure that the majority of benign and malignant nodules concentrate in the 5–10 mm range.

In this paper, the location information and the degree of malignancy of pulmonary nodules in the patient's XML

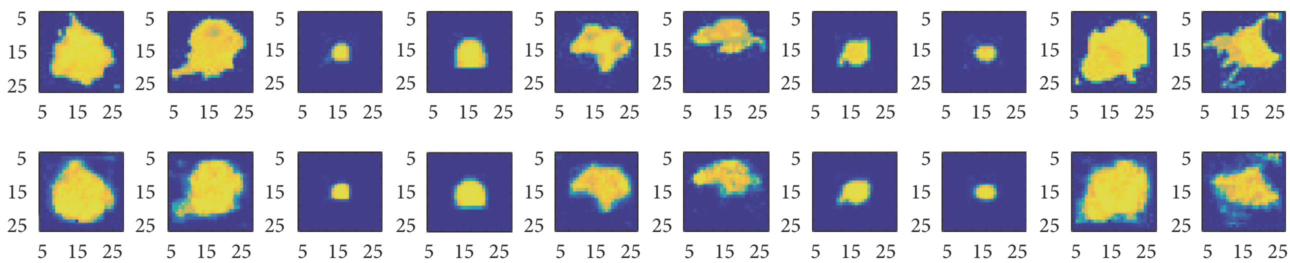


Figure 6: Autoencoder generates the pulmonary nodule image and original image.

Table 3: The structure of the SAE.

Layer	Type	Input	Output
1	Input	$28 \times 28 \times 1$	784×1
2	Fully connected	784×1	256×1
3	Fully connected	256×1	64×1
4	Fully connected	64×1	2×1
5	Softmax	2×1	Result

commentary file both can be obtained. In the XML file, four radiologists would analyze the details of the pulmonary nodules. Radiologists classify the degree of malignancy of pulmonary nodules into five categories:

(1) Highly unlikely for cancer

(2) Moderately unlikely for cancer

(3) Indeterminate likelihood

(4) Moderately suspicious for cancer

(5) Highly suspicious for cancer.

The first two categories are identified as benign. The latter two categories were identified as malignant. As a total, 9106 nodular images are obtained.

3.6. Data Augmentation. It is known that the sizes of the pulmonary nodule are different. In order to obtain the textural and size characteristics of the lung nodules, the size of the pulmonary nodules is set at 28×28 uniformly. Firstly, the image of the pulmonary nodules was obtained by binary processing, which can obtain the approximate outline of the pulmonary nodules. Then, the value of the pulmonary nodules was restored in the proceeded image to the pixels of the pulmonary nodules. Finally, noise disturbance around pulmonary nodules can be eliminated. The original images and binary images contrast in Figure 5.

A large number of positive samples and negative samples are needed to satisfy the neural network training. In this paper, the image processing operation of translation, rotation, and flip is obtained before the image was input into the neural network, which increased the sample data of the input image. Large number of sample data can effectively improve the neural network training and testing accuracy, reduce the loss function, and ultimately improve the robustness of neural networks.

Table 4: Results for all architectures.

Models	Accuracy	Sensitivity	Specificity
CNN	84.15%	83.96%	84.32%
DNN	82.37%	80.66%	83.9%
SAE	82.59%	83.96%	81.35%

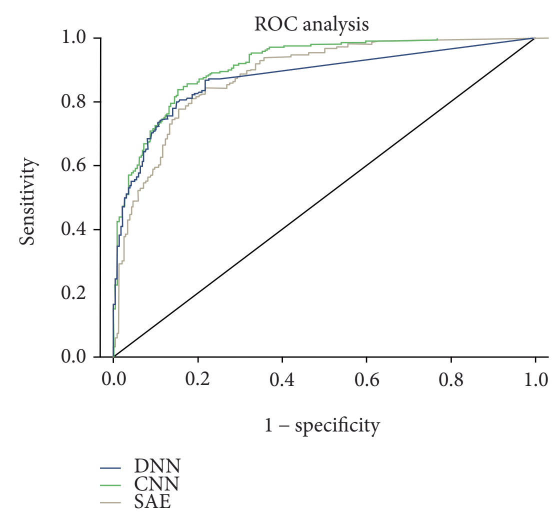


Figure 7: ROC curve of different neural networks.

4. Experiments and Results

4.1. Experiment Setup. Caffe which is a deep learning framework made with expression, speed, and modularity in mind was used in this study. A total of 4581 images of lung nodules were used in the training. Among them, 2265 cases were benign pulmonary nodules and the other one was malignant pulmonary nodules with 2311 images. 10% of the training data set is used for cross-validation, about 448 pictures. The same data set is applied to the three different kinds of network architecture.

4.1.1. Construction of the CNN. Using the network in the training stage, CNN learning rate is set to 0.01 and batch_size to 32, to get the best results. In the network, the convolution operation and the down sampling operation are carried out two times. Two convolution layers consist of 32 filters, and the kernel size is 5. The pooling layer has a kernel size of 2.

Table 5: Comparison with other papers.

Work	Database (samples)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Nascimento et al. [21]	LIDC (73)	92.78	85.64	97.89
Orozco and Villegas [22]	NBIA-ELCAP (113)	N/A	96.15	52.17
Krewer et al. [7]	LIDC-IDRI (33)	90.91	85.71	94.74
Dandil et al. [23]	Private (128)	90.63	92.30	89.47
Parveen and Kavitha [24]	Private (3278)	N/A	91.38	89.56
Kuruvilla and Gunavathi, 2014 [6]	LIDC (110)	93.30	91.40	100
Gupta and Tiwari [25]	Private (120)	90	86.66	93.33
Hua et al. [10]	LIDC (2545)	N/A	73.30	78.70
Kumar et al. [8]	LIDC (4323)	75.01	83.35	N/A
da Silva [26]	LIDC-IDRI (8296)	82.3	79.4	83.8
CNN (this paper)	LIDC-IDRI (5024)	84.15%	83.96%	84.32%
DNN (this paper)	LIDC-IDRI (5024)	82.37%	80.66%	83.9%
SAE (this paper)	LIDC-IDRI (5024)	82.59%	83.96%	81.35%

The reason of using a dropout layer is to prevent overfitting. Two fully connected layers and a softmax function is following at least.

4.1.2. Construction of the DNN. The DNN consists of a fully connected layer. The input image is a two-dimensional data input 28×28 neural network mapped into 784×1. The second layer is a fully connected layer of 512×1. The third layer is a fully connected layer of 256×1. After the third layer, there will be a dropout layer, with a parameter of 0.6, in which the unit will be hidden in 40%. The fourth layer is a fully connected layer of 64×1, whose activation function is set to ReLU.

4.1.3. Construction of the SAE. The SAE is also made up of a fully connected layer. The neurons of the autoencoder's input and output are the same; the autoencoder is equivalent to the following function:

$$H_{w,b}(x) = x, \tag{4}$$

where w and b are the weight and crankiness, respectively, in the neural network operation and x is the input parameter. The neural network is equivalent to coding the input image. Because of the problem of image classification, the hidden layer generated by the self-encoder is directly used for classification, thus canceling the decoding part of the self-encoder.

During the training, the encoder-generated stack encoding is used firstly, and then, the coding part of the stack encoding network is used to apply the initializing neural network after a certain number of training to the classification. In Figure 6, the image is the contrast between the autoencoder that generates the pulmonary nodule image and original image. It is found that the image after the encoder has made the edge of the image and the characteristics of the artifacts are not obvious. So the classification accuracy will cause some loss. The detail of the SAE is in Table 3.

4.2. Results and Analysis. As referred in Table 4, the CNN architecture has the best precision, with an accuracy of 84.15%, sensitivity of 83.96%, and specificity of 84.32%. The accuracy of the DNN is 82.37%, the sensitivity is 80.66%, and the specificity is 83.9%. The convolution neural network obtains the good result mainly because the convolution layer operation may obtain the characteristic from the shape and the texture of two different dimensions. In different convolution kernels according to different weights for different image characteristics, a convolution kernel shared parameters in the whole process of convolution, so the convolution operation compared with fully connected operation has fewer parameters. Compared with the SAE, the DNN is not good in precision and sensitivity, but it has a better effect on specificity of 83.9%. Good specificity means that more malignant lung nodules can be detected in the same data set, which may be of a greater help in the early diagnosis of pulmonary nodules. But to a certain extent, the DNN increases the number of false-positive pulmonary nodules. The SAE and DNN are consisting only of fully connected networks, but there are different ways of generating. The SAE is generated through sparsing since the encoder training; the DNN is generated through the fully connected layer directly since training.

In order to compare the performance of the neural network, the ROC curve is used in the paper. Figure 7 is the comparison of the ROC curves of the three different neural network architectures, from which we can see that the performance of the CNN is better than that of the SAE. The AUC of the CNN is 0.916, of the SAE is 0.884, and of the DNN is 0.877.

Table 5 shows some of the relevant work and the results of this comparison. In order to increase the comparability, the experiments in the paper are done in the same data set, as well as the comparison of the same parameters. By contrast, the experimental data and the results of the CNN architecture have made some progress.

5. Conclusion

In this paper, three important deep neural networks were exploited and extensively evaluated. The prediction in the classification of benign and malignant pulmonary nodules was compared in LIDC-IDRI. The experimental results suggest that the CNN archived the best performance than the DNN and SAE. The layers of the neural network in this paper are relatively small, due to the limitations of the data sets. The proposed method can be expected to improve accuracy of the other database. The method can be generalized to the design of high-performance CAD systems for other medical imaging tasks in the future.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research work was funded in part by the Tianjin Higher School Science and Technology Development Fund (20130806) and the National Natural Science Foundation of China Youth Science Fund Project (61403276).

References

[1] *Key Statistics for Lung Cancer*, https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.

[2] A.C. Society, *Cancer Facts and Figures*, 2015, http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-044552.pdf.

[3] I. R. Valente, P. C. Cortez, E. C. Neto, J. M. Soares, V. H. de Albuquerque, and J. M. Tavares, "Automatic 3D pulmonary nodule detection in CT images: a survey," *Computer Methods and Programs in Biomedicine*, vol. 124, no. 1, pp. 91–107, 2016.

[4] A. El-Baz, G. M. Beache, G. Gimel'farb et al., "Computer-aided diagnosis systems for lung cancer: challenges and methodologies review article," *International Journal of Biomedical Imaging*, vol. 2013, Article ID 942353, 46 pages, 2013.

[5] H. Chen and W. WuH. Xia, J. Du, M. Yang, and B. Ma, "Classification of pulmonary nodules using neural network ensemble," in *Advances in Neural Networks*, pp. 460–466, Springer, Guilin, China, 2011.

[6] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CT images," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 1, pp. 202–209, 2014.

[7] H. Krewer, B. Geiger, L. O. Hall et al., "Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2013*, pp. 3887–3891, IEEE, Manchester, United Kingdom, 2013.

[8] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in CT images," in *12th Conference on Computer and Robot Vision (CRV)*, pp. 133–138, IEEE, 2015.

[9] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[10] K. L. Hua, C. H. Hsu, S. C. Hidayati, W. H. Cheng, and Y. J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *OncoTargets and Therapy*, vol. 8, pp. 2015–2022, 2014.

[11] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Proceedings of 24th International Conference on Information Processing in Medical Imaging*, pp. 588–599, 2015.

[12] H. C. Shin, H. R. Roth, M. Gao et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[13] LIDC-IDRI, https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.

[14] M. Browne and S. S. Ghidary, "Convolutional neural networks for image processing: an application in robot vision," in *AI 2003: Advances in Artificial Intelligence*, pp. 641–652, 2003.

[15] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 253–256, IEEE, 2010.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from over fitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[19] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[20] Y. Jia, E. Shelhamer, J. Donahue et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, 2014.

[21] L. B. Nascimento, A. C. de Paiva, and A. C. Silva, "Lung nodules classification in CT images using Shannon and Simpson diversity indices and SVM," in *Machine Learning and Data Mining in Pattern Recognition*, pp. 454–466, 2012.

[22] H. M. Orozco and O. O. V. Villegas, "Lung nodule classification in CT thorax images using support vector machines," in *12th Mexican International Conference on Artificial Intelligence*, pp. 277–283, IEEE, 2013.

[23] E. Dandıl, M. Çakiroğlu, Z. Ekşi, M. Özkan, Ö. K. Kurt, and A. Canan, "Artificial neural network-based classification system for lung nodules on computed tomography scans," in *6th International Conference of Soft Computing and Pattern Recognition (soCPar)*, pp. 382–386, IEEE, 2014.

[24] S. S. Parveen and C. Kavitha, "Classification of lung cancer nodules using SVM kernels," *International Journal of Computer Applications*, vol. 95, p. 25, 2014.

[25] B. Gupta and S. Tiwari, "Lung cancer detection using curvelet transform and neural network," *International Journal of Computer Applications*, vol. 86, p. 1, 2014.

[26] G. L. F. da Silva, A. C. Silva, A. C. de Paiva, and M. Gattass, *Classification of Malignancy of Lung Nodules in CT Images Using Convolutional Neural Network*.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 9856058, 9 pages
https://doi.org/10.1155/2017/9856058

Research Article

Low-Rank and Sparse Decomposition Model for Accelerating Dynamic MRI Reconstruction

Junbo Chen,[1,2] Shouyin Liu,[1] and Min Huang[2,3]

[1] *College of Physical Science and Technology, Central China Normal University, Wuhan 430079, Hubei, China*
[2] *Key Laboratory of Cognitive Science of State Ethnic Affairs Commission, South-Central University for Nationalities, Wuhan 430074, Hubei, China*
[3] *Hubei Key Laboboratory of Medical Information Analysis & Tumor Diagnosis and Treatment, Wuhan 430074, Hubei, China*

Correspondence should be addressed to Shouyin Liu; syliu@phy.ccnu.edu.cn

Received 10 March 2017; Accepted 17 May 2017; Published 8 August 2017

Academic Editor: Feng-Huei Lin

The reconstruction of dynamic magnetic resonance imaging (dMRI) from partially sampled *k*-space data has to deal with a trade-off between the spatial resolution and temporal resolution. In this paper, a low-rank and sparse decomposition model is introduced to resolve this issue, which is formulated as an inverse problem regularized by robust principal component analysis (RPCA). The inverse problem can be solved by convex optimization method. We propose a scalable and fast algorithm based on the inexact augmented Lagrange multipliers (IALM) to carry out the convex optimization. The experimental results demonstrate that our proposed algorithm can achieve superior reconstruction quality and faster reconstruction speed in cardiac cine image compared to existing state-of-art reconstruction methods.

1. Introduction

Dynamic MRI (magnetic resonance imaging), an essential medical imaging technique, allows noninvasiveness, nonionization visualization, and analysis of anatomical and functional changes of internal body structure through time. However, MRI sampling speed is relatively slow due to the need of physical and physiological conditions such as nuclear relaxation and peripheral nerve stimulation [1]. One way for accelerating MRI is to reconstruct high-resolution images from undersampled *k*-space data. However, such undersampling violates the Nyquist criterion and often results in aliasing artifacts if the traditional linear reconstruction is directly applied.

To address this issue, there are so much research efforts to accelerate MRI acquisition process using hardware and software [2–4]. Among them, compressed sensing (CS) has been proved to be able to increase imaging speed and efficiency in MRI application [5–7]. The CS theory requires image sparsity and incoherence between the acquisition space and representation space [8]. Fortunately, the MR image sequence often provides redundant information in both spatial and temporal domains, which presents favorable conditions for the application of CS. In addition, the idea is easily extended to the reconstruction of dynamic MRI (dMRI) images due to extensive spatio-temporal correlations that result in sparser representations. The *k-t* FOCUSS is a successful method, which imposes a sparsity constraint in the temporal transform domain by using the FOCUSS algorithm [9], and extends the FOCUSS technique with motion estimation and compensation to compressed sensing framework for cardiac cine MRI. But the limitation of the prediction schemes on sparsifying the residual signal sets back the further improvement when the motion is aperiodic.

Recently, researchers have made great efforts to exploit the low-rank property of matrices instead of simply sparsity of vectors. Lingala et al. proposed a *k-t* SLR algorithm that exploited the low rank prior and global sparsity in

Karhunen-Louve Transform (KLT) domain for MRI reconstruction [10]. However, the algorithm failed to take into account the structural sparsity of the MRI image, and the limitation held back the further improvement. Some studies presented patch-based dictionary learning techniques for dMRI reconstruction [11, 12]. However, a major challenge in learning sparse dictionary is that such patch-based learning cannot be effectively employed for dMRI reconstruction. Because the size of dMRI sequence is large, it is inefficient to learn dictionaries for such large datasets [13]. Even though we take no account of computational limitations, it is not practical to acquire such huge dMRI training sequences for learning sparsifying dictionaries. Currently, robust principal component analysis (RPCA) has been used in recovering dynamic images to explore the low-rank structure of data [14, 15]. The RPCA decomposes the data in low rank and sparse components, where the low rank component models the temporally correlated background information and the sparse component represents the dynamic information. *k-t* RPCA [16], a method developed for dMRI, uses the low-rank plus sparse decomposition prior to reconstructing dynamic MRI from part of the *k*-space measurements. In this method [16], the image reconstruction is regularized by a low-rank plus sparse prior, where the Fourier transform is used as the sparsifying transform and the alternating direction methods of multipliers (ADMM) is applied to solve the minimization problem in the temporal direction. The shortcoming of *k-t* RPCA is that the results of reconstructed image are easily affected by the noise, since the noise will generally be represented by highly sparse coefficients during the sparsifying transform.

In this paper, aims to the shortcoming of *k-t* RPCA, we propose an efficient numerical algorithm based on inexact augmented Lagrangian method (IALM) instead of ADMM to solve the optimization problem and accelerate the dMRI reconstruction. The experimental results demonstrate that our proposed algorithm can achieve more satisfactory reconstruction performance and faster reconstruction speed in given cardiac cine sets.

2. Theory Background

The dynamic MRI data acquisition in the *k-t* space can be expressed as follows:

$$y(k,t) = \int x(r,t)\exp(-2\pi jk\cdot r)dr + n(k,t), \quad (1)$$

where $y(k,t)$ represents the measured *k-t* space signal, $x(r,t)$ denotes the desired dynamic image series, and $n(k,t)$ is the measured noise, which can be reasonably modeled by an additive white Gaussian distribution [16, 17].

In this paper, the solution of this problem is to find the closest representation of the MR image $x(r,t)$ from undersampled measurement $y(k,t)$. Since the *k-t* space is partially sampled, (1) is converted to an inverse problem and can be rewritten as a vector [18].

$$\mathbf{Y} = \mathbf{R}F\mathbf{X} + \mathbf{n}, \quad (2)$$

where $\mathbf{Y} = [y_1|\cdots|y_T]$, $\mathbf{X} = [x_1|\cdots|x_T]$, $\mathbf{n} = [n_1|\cdots|n_T]$, T is the total number of frames, F is the Fourier transform operator, and the measurement matrix $\mathbf{R}$ is the undersampled mask applied on the *k*-space.

2.1. CS-Based MR Image Reconstruction. The CS approach [5, 19] was proposed to reconstruct the MR image $\mathbf{X}$ from the partially sampled *k*-space data $\mathbf{Y}$ by exploiting the sparsity transform and convex optimization algorithms. The problem will be solved if we can find the sparsest vector satisfying (2),

$$\min_X \|D\mathbf{X}\|_0 \quad \text{s.t.} \quad \|\mathbf{Y} - \mathbf{R}F\mathbf{X}\|_2 \le \varepsilon, \quad (3)$$

where $\|\cdot\|_0$ is l_0-norm, counting the number of nonzero entries in the vector, D is the sparsifying transform or dictionary, and ε is a small constant. Unfortunately, (3) is NP-hard problem, which needs to be solved by a brute force search. The CS theory [8] proves that the convex relaxation approach referred to as l_1 minimization can be replaced with the l_0-norm in (3),

$$\min_X \|D\mathbf{X}\|_1 \quad \text{s.t.} \quad \|\mathbf{Y} - \mathbf{R}F\mathbf{X}\|_2 \le \varepsilon, \quad (4)$$

where $\|\cdot\|_1$ is l_1-norm, meaning the sum of absolute values of the vector.

2.2. Low-Rank and Sparse Decomposition Model for MR Image Reconstruction. CS-based techniques that exploit sparsity of the image in the transform domain have been successfully used for MR image reconstruction. However, the performance of CS is primarily dependent on the specific dictionary or sparsifying operator, which limits the maximum achievable acceleration rate. Therefore, some researchers tried to investigate a few new approaches to reconstruct MR image [20–24]. In those methods, low-rank matrix recovery is a popular technique in medical image processing.

The basic assumption is the same as [18], that is, the image $\mathbf{X}$ is simultaneously sparse (in a transform domain) and low rank. The problem is to recover $\mathbf{X}$, given fewer *k*-space samples $\mathbf{Y}$ than the number of elements in the matrix. We assume that the approximate rank of the matrix is r and the size of single frame image is $M \times N$. When the matrix $\mathbf{X}$ is low rank, which has only $r(M+N-r)$ degrees of freedom instead of MN, it is possible to recover the matrix $\mathbf{X}$ from lesser number of samples by solving the rank minimization problem,

$$\min_{\mathbf{X}} \operatorname{rank}(\mathbf{X}) \quad \text{s.t.} \quad \|\mathbf{Y} - M(\mathbf{X})\|_2 \le \varepsilon. \quad (5)$$

However, the rank minimization problem, that is, solving (5), is combinatorial and known to be NP-hard [25]. Therefore, convex relaxation is often used to make the minimization tractable.

$$\min_{\mathbf{X}} \|\mathbf{X}\|_* \quad \text{s.t.} \quad \|\mathbf{Y} - M(\mathbf{X})\|_2 \le \varepsilon, \quad (6)$$

where M denotes any linear operator and $\|\mathbf{X}\|_*$ is the nuclear norm, which is defined as

$$\|\mathbf{X}\|_* = \sum_{i=1}^{r} \sigma_i, \tag{7}$$

where $\sigma_1, \sigma_2, \ldots, \sigma_r$ are the singular values of $\mathbf{X}$ and r is the rank of $\mathbf{X}$.

To recover $\mathbf{X}$ from the given $\mathbf{Y}$, $\mathbf{X}$ can be decomposed into a superposition of a low-rank matrix $\mathbf{A}$ and a sparse matrix $\mathbf{E}$.

$$\mathbf{X} = \mathbf{A} + \mathbf{E}. \tag{8}$$

$\mathbf{X}$ is recovered as the solution of the following optimization:

$$\min_{A,E} \|\mathbf{A}\|_* + \gamma\|\mathbf{E}\|_1 \quad \text{s.t.} \quad \|\mathbf{Y} - M(\mathbf{A}+\mathbf{E})\|_2 \le \varepsilon, \tag{9}$$

where low-rank matrix $\mathbf{A}$ has few nonzero singular values and represents the background component, sparse matrix $\mathbf{E}$ has few nonzero entries and corresponds to the changes, and γ is a tuning parameter that balances the contribution of the l_1-norm relative to the nuclear norm.

3. The Proposed Method

In principal component pursuit (PCP) model [26], to solve (9) can be posed as an optimization problem by using regularization rather than strict constraints [15]. Hence, (9) can be converted as

$$\min_{L,S} \|\mathbf{Y} - M(\mathbf{A}+\mathbf{E})\|_F^2 + \lambda_L\|\mathbf{A}\|_* + \lambda_S\|T\mathbf{E}\|_1, \tag{10}$$

where the parameters λ_L and λ_S trade off data consistency and T is a sparse transform basis.

Equation (10) is a RPCA problem that involves minimizing a combination of the nuclear norm and l_1-norm. Otazo et al. Study [15] adopted the iterative thresholding scheme to solve (10); however, the iterative thresholding technique converges slowly. So, we presented an inexact augmented Lagrange multipliers (IALM) algorithm to solve the RPCA problem [27]. According to the constraint conditions of (6),

$$M^H y = \mathbf{A}^{(n)} + \mathbf{E}^{(n)} = \mathbf{X}^{(n)}, \tag{11}$$

where M^H is a dual operator, $\mathbf{X}^{(n)}$ contains the measurement noise, and $\mathbf{A}^{(n)}$ and $\mathbf{E}^{(n)}$ are low-rank element and sparse element, respectively. We applied IALM method to solve the following optimization problem:

$$\min_{L,S} \left\|\mathbf{A}^{(n)}\right\|_* + \lambda\left\|T\mathbf{E}^{(n)}\right\|_1 + \langle \mathcal{L}, M^H y - \mathbf{A}^{(n)} - \mathbf{E}^{(n)} \rangle + \frac{\mu}{2}\left\|M^H y - \mathbf{A}^{(n)} - \mathbf{E}^{(n)}\right\|_F^2, \tag{12}$$

where $\mathcal{L}$ is a Lagrange multiplier to remove the equality constraint and μ is a small positive scalar. The condition $\sum_{k=1}^{+\infty} \mu_k^{-1} = +\infty$ implies that μ_k cannot grow too fast. The IALM method for solving the RPCA problem can be described as Algorithm 1.

For Algorithm 1, if $\{\mu_k\}$ is nondecreasing and $\sum_{k=1}^{+\infty} \mu_k^{-1} = +\infty$, then $(\mathbf{A}_k, \mathbf{E}_k)$ converges to an optimal solution $(\mathbf{A}^*, \mathbf{E}^*)$ for the RPCA problem. The advantage of unbounded $\{\mu_k\}$ is that the feasibility condition $\mathbf{A}_k + \mathbf{E}_k = \mathbf{X}$ can be approached more quickly because $\mathbf{X} - \mathbf{A}_k - \mathbf{E}_k = (\mathcal{L}_k - \mathcal{L}_{k-1})/\mu_{k-1}$ and $\{\mathcal{L}_k\}$ are bounded. In Algorithm 1, the singular value thresholding (SVT) operator [28] is defined as

Input: Multicoil undersampled *k-t* data y; space-time multicoil encoding operator M; sparsifying transform T; weight parameter λ.
Initialization: Set $\mathcal{L}_0 = 0$; $\boldsymbol{E}_0 = 0$; $k = 0$; $\mu_0 > 0$; $\boldsymbol{X}_0 = M^H y$; $\rho > 0$; $iter = 0$.
Loop: Repeat until convergence
  Update $\boldsymbol{A}$: $(U, S, V) = svd(\boldsymbol{X}_k - \boldsymbol{E}_k + \mu_k^{-1}\mathcal{L}_k)$;
    $\boldsymbol{A}_{k+1} = U\Lambda_{\mu_k^{-1}}[S]V^T$.
  Update $\boldsymbol{E}$: $\boldsymbol{E}_{k+1} = \Lambda_{\lambda\mu_k^{-1}}[\boldsymbol{X}_k - \boldsymbol{A}_{k+1} + \mu_k^{-1}\mathcal{L}_k]$.
  Update $\mathcal{L}$: $\mathcal{L}_{k+1} = \mathcal{L}_k + \mu_k(\boldsymbol{X}_k - \boldsymbol{A}_{k+1} - \boldsymbol{E}_{k+1})$.
  Update $\boldsymbol{\mu}_{k+1}$: $\mu_{k+1} = \rho\mu_k$.
  Update $\boldsymbol{X}_{k+1}$: $\boldsymbol{X}_{k+1} = \mathbf{A}_{k+1} + \boldsymbol{E}_{k+1} - M^H(M(\mathbf{A}_{k+1} + \boldsymbol{E}_{k+1}) - y)$.
Update $\boldsymbol{k}$, ***iter***: set $k \leftarrow k+1$; $iter \leftarrow iter + 1$.
End Loop.
  Output: $(\boldsymbol{A}_k, \boldsymbol{E}_k)$

Algorithm 1: The proposed algorithm.

$$SVT_\lambda(\mathbf{D}) = U\Lambda_\lambda(\Sigma)V^H, \tag{13}$$

where $\mathbf{D} = U\Sigma V^H$ is any singular value decomposition of $\mathbf{D}$. $\Lambda_\lambda(\Sigma)$ is a soft-thresholding operator, which can be defined as

$$\Lambda_\lambda(x) = \frac{x}{|x|}\max(|x| - \lambda, 0). \tag{14}$$

4. Experimental Results and Discussion

Experiments were run in MATLAB V7.14.0 (R2012a) with the computing environment being an Intel Core i7-2640 M CPU, 4.0 GB memory, and a 64-bit Win7 operating system. The proposed algorithm was validated by experiments using two cardiac cine sets. The first dataset was obtained from Bio Imaging and Signal Processing Lab (http://bispl.weebly.com/), which contains $n_t = 25$ temporal frames of size $n_x = n_y = 256$ with a $345 \times 270\,\text{mm}^2$ field of view (FOV) and 10 mm slice thickness. The second dataset was acquired from the website of Dr. Caballero (http://www.doc.ic.ac.uk/~jc1006/index.html), which was introduced by Caballero et al. [12] and the relevant imaging parameters were as follows: the image matrix size $= 256 \times 256$ $(n_x \times n_y)$, the number of temporal frame $= 30\,(n_t)$, FOV $= 320 \times 320\,\text{mm}^2$, and slice thickness =10 mm. Two widely used sampling trajectories, Cartesian and radial undersampling strategies, were exploited for the acquisition of the MR data set in the k-space domain. Figure 1 shows the sampling masks used in the study and their effect on the magnitude of a temporal frame.

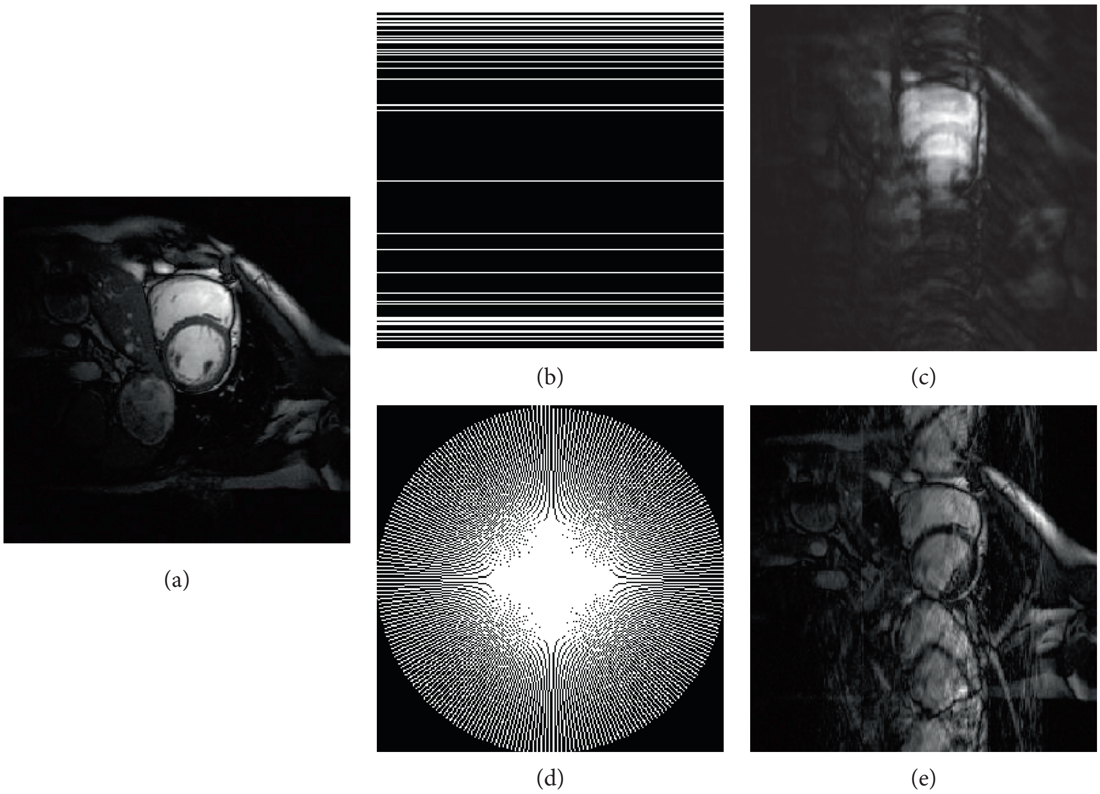


Figure 1: Example of two undersampling masks and their effect on reconstruction image. (a) A magnitude temporal frame from one of the cardiac cine datasets. (b) The Cartesian undersampling mask for a magnitude temporal frame. (c) The zero-filled inverse Fourier transform reconstruction image. (d) The pseudo-radial sampling acquisition for a magnitude temporal frame. (e) The zero-filled inverse Fourier transform reconstruction image.

We compared the proposed method against *k-t* SLR [10] and *k-t* RPCA [16] in reconstruction accuracy and reconstruction speed. Quantitative image quality assessment was performed by using the metrics of peak signal to noise ratio (PSNR) and structural similarity index (SSIM) [29]. The PSNR is used to evaluate the difference of reconstruction image and the full-sampled image, which can be defined as

$$\mathrm{PSNR} = -10\log_{10}\left(\frac{\left\|\widehat{\mathbf{X}}-\mathbf{X}\right\|_F^2}{\left\|\mathbf{X}\right\|_F^2}\right), \tag{15}$$

where $\widehat{\mathbf{X}}$ is a reconstruction image and $\mathbf{X}$ represents the full-sampled image.

The SSIM is a new method for measuring the similarity between the reconstructed image and the fully sampled image. We adopted the SSIM to measure the difference of reconstruction image and the fully sampled image at each time frame $\{x_n\}_{n=1}^{n_t}$; the SSIM index between the reconstructed image x_{Rec} and the fully sampled image x_F at one same frame is evaluated as

$$\mathrm{SSIM}(x,y) = \frac{\left(2\mu_{x_R}\mu_{x_F}+c_1\right)\left(2\sigma_{x_R x_F}+c_2\right)}{\left(\mu_{x_R}^2+\mu_{x_F}^2+c_1\right)\left(\sigma_{x_R}^2+\sigma_{x_F}^2+c_2\right)}, \tag{16}$$

where μ_{x_R} and μ_{x_F} are the mean intensity of the reconstructed image x_{Rec} and the fully sampled image x_F, σ_{x_R} and σ_{x_F} are the standard deviation of image x_{Rec} and x_F, $\sigma_{x_R x_F}$ is the covariance of x_{Rec} and x_F, $c_1 = (K_1 L)^2$ and $c_2 = (K_2 L)^2$ are constants where L is the dynamic range, 255 for 8-bit grayscale images. $K_1 = 0.01$ and $K_2 = 0.03$ are parameter values suggested by Wang et al. [29].

The *k-t* SLR uses a combination of TV and nonconvex Schatten p-norms with $p = 0.01$; some parameters are selected based on the suggested values in the public software package (penalty parameters $\beta_1 = 10^{-9}$ for Schatten and $\beta_2 = 10^{-2}$ for TV norms, maximum number of 50 inner and 9 outer iterations). In *k-t* RPCA, two regularization parameters are $\mu = 200$ and $\rho = 1.5$ for the regularization and decomposition, respectively.

Similarly, the method requires the specification of three parameters λ, ρ, and μ. We set $\mu_0 = 1.5/\|\mathbf{X}\|_2$ and $\rho = 1.2$. We may take $\|\mathbf{X}-\mathbf{A}_k-\mathbf{E}_k\|_F/\|\mathbf{X}\|_F < 10^{-7}$ as the stopping criteria for Algorithm 1. We chose a fixed weight parameter $\lambda = \max(n_x * n_y, n_t)^{-1/2}$ by the suggestion of the authors of [16]. The proposed algorithm was verified by experiments using the fully sampled cardiac cines (two mentioned datasets above) with two different sampling trajectories.

For simulating the acceleration of the *k*-space, the fully sampled *k*-space data was artificially subsampled by using

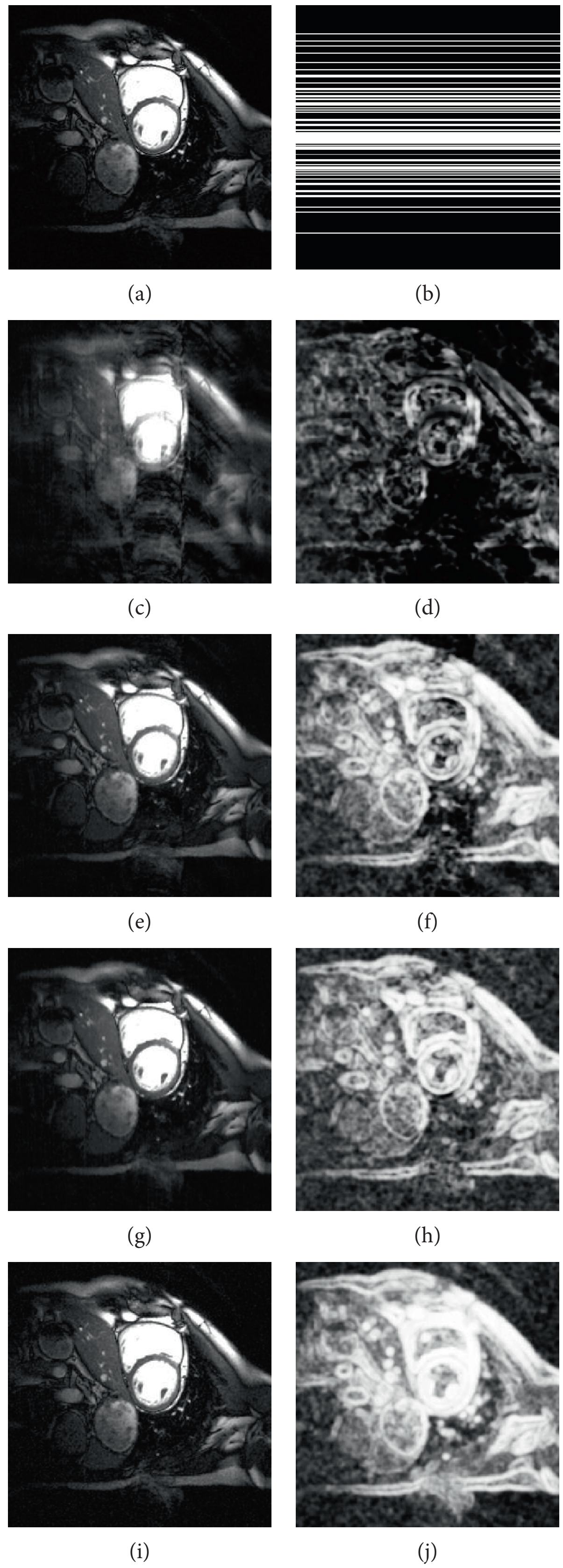


Figure 2: Comparison of the reconstruction results with different methods on the first cardiac dataset. The acceleration rate is 4 or sampling ratio is about 0.25. Fully sampled image (a) and undersampling mask (b), undersampled by zero-filled directly (c), reconstructions using *k-t* RPCA (e), *k-t* SLR (g), and proposed method (i) with their respective residuals (d, f, h, j).

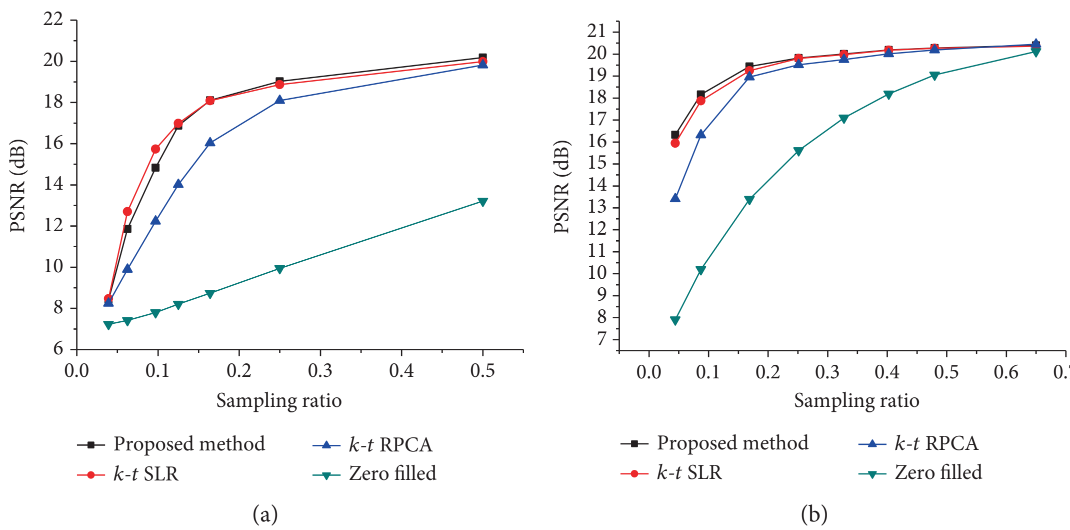


FIGURE 3: PSNR performance of different reconstructions evaluated versus the sample ratio for the first cardiac MRI dataset with Cartesian sampling (a) and radial sampling (b).

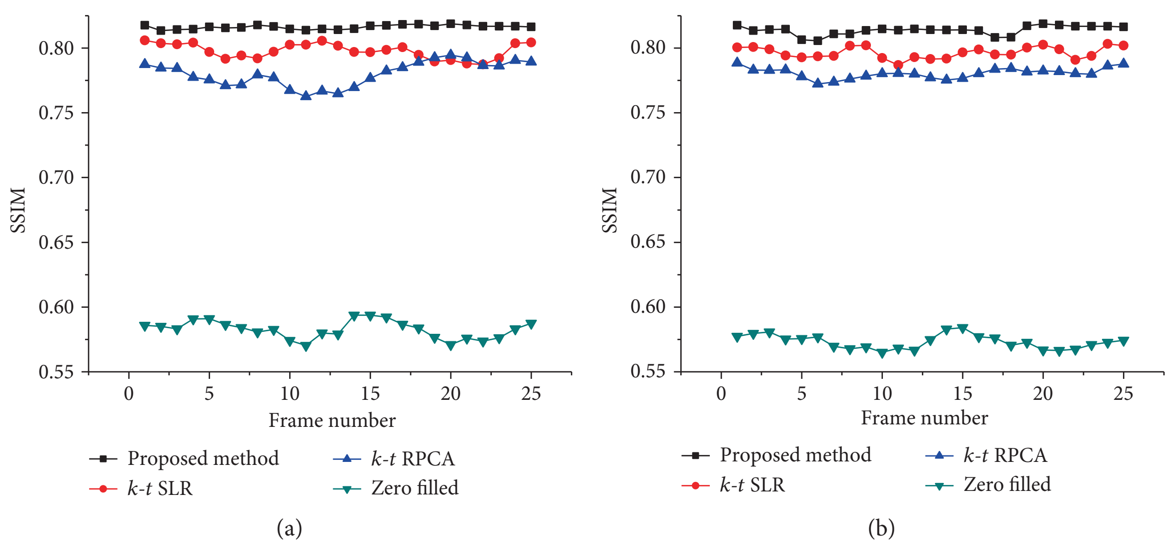


FIGURE 4: SSIM performance of different reconstructions versus each time frame for the first cardiac MRI dataset at acceleration factor 6 with Cartesian sampling (a) and radial sampling (b).

variable density (sampling factor) random sampling. To test the robustness of the proposed method, the *k*-space data of the two datasets are corrupted with additional complex Gaussian white noise with fixed standard deviation $\sigma = 15$. Firstly, this method was tested on the first cardiac dataset by using different sampling models with variable sampling ratio. A comparison of the visual quality was showed in Figure 2, which compares the reconstruction results between the proposed method (Algorithm 1) and the other methods. The acceleration factor is approximately 4 (about 25% of acquired samples) for Cartesian sampling masks. Figure 3 shows the PSNR of the reconstructed results for Cartesian sampling and pseudo-radial sampling as a function of sampling factor. It can be seen that the performance of the proposed method outperforms the other two methods with pseudo-radial sampling acquisition. But the performance at lower sampling ratio is slightly lower than the *k*-*t* SLR method with Cartesian sampling. Additionally, SSIM at each time frame is shown in Figure 4 for both Cartesian and radial sampling at the same sampling factor (an acceleration factor of approximately 6 with about 16.4% of acquired samples). The experimental results reveal that the proposed method achieved a superior reconstruction result in terms of SSIM and hence the advantage of our method is more relatively evident when pseudo-radial sampling is used instead of Cartesian sampling.

Moreover, we tested our proposed method on the second cardiac dataset by using same experimental method. Figure 5 provides the visual evaluations for radial sampling with an acceleration rate of 4 (about 25% of acquired samples). Quantitative results (PSNR performance) are reported for Cartesian sampling and pseudo-radial sampling in Figure 6.

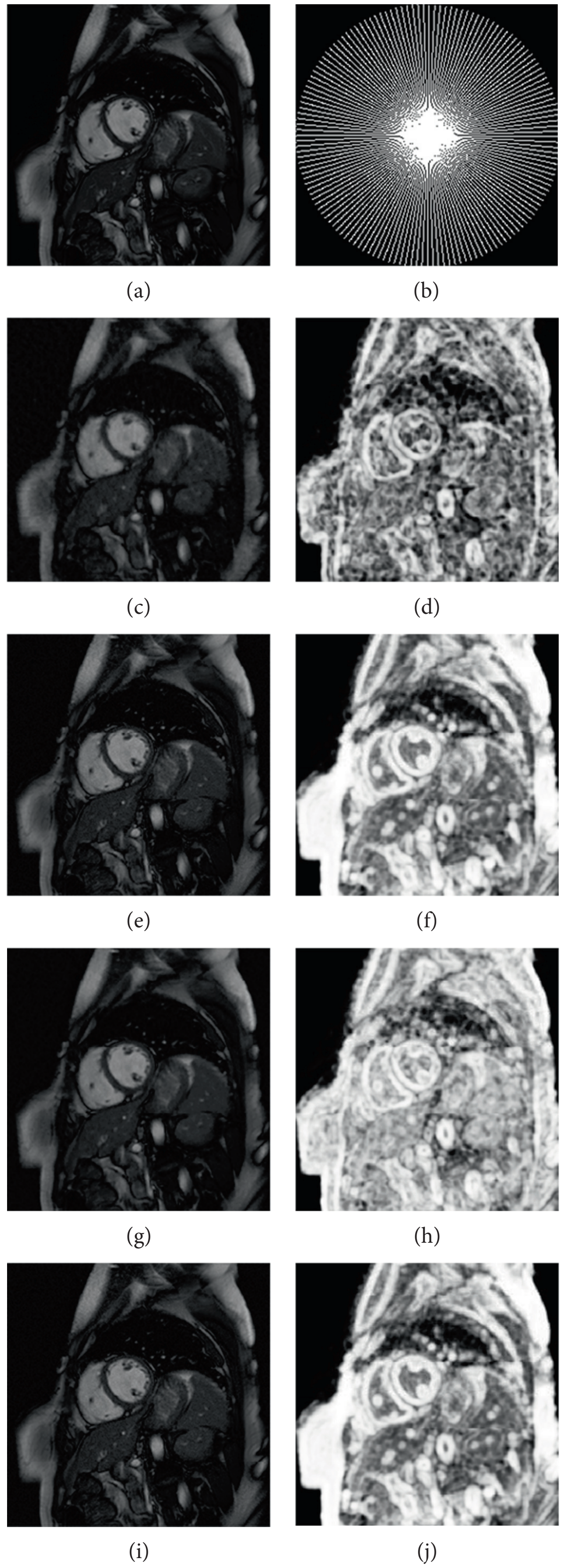


Figure 5: Comparison of the reconstruction results with different methods on the second cardiac dataset. The number of radial trajectory is 74 and sampling ratio is about 0.25. Fully sampled image (a) and undersampling mask (b), undersampled by zero-filled directly (c), reconstructions using *k-t* RPCA (e), *k-t* SLR (g), and proposed method (i) with their respective residuals (d, f, h, j).

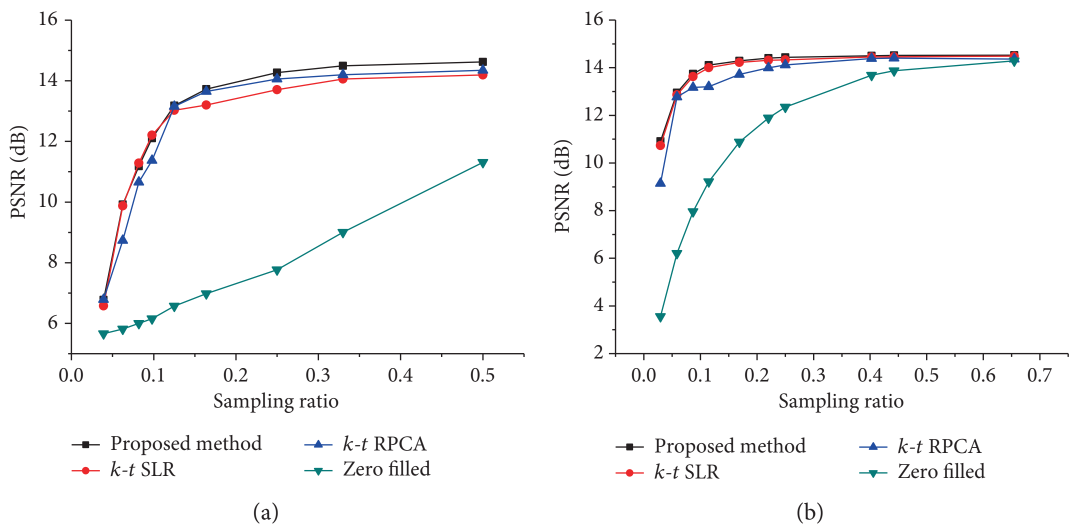


FIGURE 6: PSNR performance of different reconstructions evaluated versus the sample ratio for the second cardiac MRI dataset with Cartesian sampling (a) and radial sampling (b).

TABLE 1: Average computational time for cardiac datasets in complete temporal frames (seconds).

Method	Cardiac dataset 1 $(256 \times 256 \times 25)$		Cardiac dataset 2 $(256 \times 256 \times 30)$	
	Cartesian sampling	Pseudo-radial sampling	Cartesian sampling	Pseudo-radial sampling
k-t SLR	1676.7	1391.5	1456.3	1515.7
k-t RPCA	610.3	766.6	423.3	681.2
Proposed method	131.2	124.6	116.2	130.6

It is observed that the performance of reconstructions by using the two sampling models is similar to Figure 3.

Figures 3 and 6 indicate that both the proposed and the other two methods are effective to the choice of Cartesian sampling with higher sampling ratios. However, the choice of pseudo-radial sampling ensures that the greater performance can be obtained at lower sampling ratios. Moreover, it can be acquired that the proposed method is more robust in a cardiac cycle.

We also evaluated the execution time of the three methods by using different sampling models with variable sampling factors on different datasets. Table 1 shows the average computational time for reconstructing the cardiac MRI images in complete temporal frames. From the Table 1, it can be known that our method is faster than the other two methods, and it is more potential for online dMRI reconstruction.

5. Conclusion

In this paper, we proposed a scalable and fast algorithm (IALM) for solving RPCA optimization problem to recover dMRI sequence from highly undersampled k-space data. Our proposed algorithm has a generalized formulation capability of separating dynamic MR data into low-rank component and sparse component. And this algorithm reconstructs and separates simultaneously dynamic MR data from partial measurement. Experiments on cardiac datasets have validated the efficiency and effectiveness compared to the state-of-the-art methods.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work was supported by the National Nature Science Foundation of China (81271659), the China Postdoctoral Science Foundation (2014 M552346), and the Special Fund for Basic Scientific Research of Central College, South-Central University for Nationalities (nos. CZP17033 and CZP17014).

References

[1] Z. Lai, X. Qu, Y. Liu et al., "Image reconstruction of compressed sensing MRI using graph-based redundant wavelet transform," *Medical Image Analysis*, vol. 27, pp. 93–104, 2016.

[2] D. K. Sodickson and W. J. Manning, "Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays," *Magnetic Resonance in Medicine*, vol. 38, no. 4, pp. 591–603, 1997.

[3] D. J. Larkman and R. G. Nunes, "Parallel magnetic resonance imaging," *Physics in Medicine and Biology*, vol. 52, no. 7, pp. R15–R55, 2007.

[4] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: sensitivity encoding for fast MRI," *Magnetic Resonance in Medicine*, vol. 42, no. 5, pp. 952–962, 1999.

[5] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: the application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[6] S. Ravishankar and Y. Bresler, "MR image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE Transactions on Medical Imaging*, vol. 30, no. 5, pp. 1028–1041, 2011.

[7] Y. Huang, J. Paisley, Q. Lin, X. Ding, X. Fu, and X. P. Zhang, "Bayesian nonparametric dictionary learning for compressed sensing MRI," *IEEE Transactions on Imaging Processing*, vol. 23, no. 12, pp. 5007–5019, 2014.

[8] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "k-t FOCUSS: a general compressed sensing framework for high resolution dynamic MRI," *Magnetic Resonance in Medicine*, vol. 61, no. 1, pp. 103–116, 2009.

[10] S. G. Lingala, Y. Hu, E. DiBella, and M. Jacob, "Accelerated dynamic MRI exploiting sparsity and low-rank structure: k-t SLR," *IEEE Transactions on Medical Imaging*, vol. 30, no. 5, pp. 1024–1054, 2011.

[11] S. P. Awate and E. V. R. DiBella, "Spatiotemporal dictionary learning for undersampled dynamic MRI reconstruction via joint frame-based and dictionary-based sparsity," in *IEEE International Symposium on Biomedical Imaging*, pp. 318–321, Barcelona, 2012.

[12] J. Caballero, A. N. Price, D. Rueckert, and J. V. Hajnal, "Dictionary learning and time sparsity for dynamic MR data reconstruction," *IEEE Transactions on Medical Imaging*, vol. 33, no. 4, pp. 979–994, 2014.

[13] A. Majumdar and R. K. Ward, "Learning the sparsity basis in low-rank plus sparse model for dynamic MRI reconstruction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 778–782, South Brisbane, QLD, 2015.

[14] A. Majumdar and R. K. Ward, "Exploiting rank deficiency and transform domain sparsity for MR image reconstruction," *Magnetic Resonance Imaging*, vol. 30, no. 1, pp. 9–18, 2012.

[15] R. Otazo, E. Candès, and D. K. Sodickson, "Low-rank and sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components," *Magnetic Resonance in Medicine*, vol. 73, no. 3, pp. 1125–1136, 2015.

[16] B. Trémoulhéac, N. Dikaios, D. Atkinson, and S. R. Arridge, "Dynamic MR image reconstruction-separation from undersampled (k,t)-space via low-rank plus sparse prior," *IEEE Transactions on Medical Imaging*, vol. 33, no. 8, pp. 1689–1701, 2014.

[17] A. Majumdar, R. K. Ward, and T. Aboulnasr, "Non-convex algorithm for sparse and low-rank recovery: application to dynamic MRI reconstruction," *Magnetic Resonance Imaging*, vol. 31, no. 3, pp. 448–455, 2013.

[18] A. Majumdar, "Improved dynamic MRI reconstruction by exploiting sparsity and rank-deficiency," *Magnetic Resonance Imaging*, vol. 31, no. 5, pp. 789–795, 2013.

[19] U. Gamper, P. Boesiger, and S. Kozerke, "Compressed sensing in dynamic MRI," *Magnetic Resonance in Medicine*, vol. 59, no. 2, pp. 365–373, 2008.

[20] J. P. Haldar and Z. P. Liang, "Spatiotemporal imaging with partially separable functions: a matrix recovery approach," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 716–719, Rotterdam, 2010.

[21] Y. Hu, S. G. Lingala, and M. Jacob, "A fast majorize-minimize algorithm for the recovery of sparse and low-rank matrices," *IEEE Transactions on Imaging Processing*, vol. 21, no. 2, pp. 742–753, 2012.

[22] B. Zhao, J. P. Haldar, A. G. Christodoulou, and Z. P. Liang, "Image reconstruction from highly undersampled-space data with joint partial separability and sparsity constraints," *IEEE Transactions on Medical Imaging*, vol. 31, no. 9, pp. 1809–1820, 2012.

[23] S. G. Lingala and M. Jacob, "Blind compressive sensing dynamic MRI," *IEEE Transactions on Medical Imaging*, vol. 32, no. 6, pp. 1132–1145, 2013.

[24] R. M. Lebel, J. Jones, J. C. Ferre, M. Law, and K. S. Nayak, "Highly accelerated dynamic contrast enhanced imaging," *Magnetic Resonance in Medicine*, vol. 71, no. 2, pp. 635–644, 2014.

[25] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computation Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[26] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in *IEEE International Symposium on Information Theory*, pp. 1518–1522, Austin, TX, 2010.

[27] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," http://arxiv.org/abs/1009.5055.

[28] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transaction on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 4037190, 9 pages
https://doi.org/10.1155/2017/4037190

Research Article

A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images

David Vázquez,[1,2] Jorge Bernal,[1] F. Javier Sánchez,[1] Gloria Fernández-Esparrach,[3] Antonio M. López,[1,2] Adriana Romero,[2] Michal Drozdzal,[4,5] and Aaron Courville[2]

[1]*Computer Vision Center, Computer Science Department, Universitat Autonoma de Barcelona, Barcelona, Spain*
[2]*Montreal Institute for Learning Algorithms, Université de Montréal, Montreal, QC, Canada*
[3]*Endoscopy Unit, Gastroenterology Service, CIBERHED, IDIBAPS, Hospital Clinic, Universidad de Barcelona, Barcelona, Spain*
[4]*École Polytechnique de Montréal, Montréal, QC, Canada*
[5]*Imagia Inc., Montréal, QC, Canada*

Correspondence should be addressed to David Vázquez; dvazquez@cvc.uab.es

Received 24 February 2017; Accepted 22 May 2017; Published 26 July 2017

Academic Editor: Junfeng Gao

Colorectal cancer (CRC) is the third cause of cancer death worldwide. Currently, the standard approach to reduce CRC-related mortality is to perform regular screening in search for polyps and colonoscopy is the screening tool of choice. The main limitations of this screening procedure are polyp miss rate and the inability to perform visual assessment of polyp malignancy. These drawbacks can be reduced by designing decision support systems (DSS) aiming to help clinicians in the different stages of the procedure by providing endoluminal scene segmentation. Thus, in this paper, we introduce an extended benchmark of colonoscopy image segmentation, with the hope of establishing a new strong benchmark for colonoscopy image analysis research. The proposed dataset consists of 4 relevant classes to inspect the endoluminal scene, targeting different clinical needs. Together with the dataset and taking advantage of advances in semantic segmentation literature, we provide new baselines by training standard fully convolutional networks (FCNs). We perform a comparative study to show that FCNs significantly outperform, without any further postprocessing, prior results in endoluminal scene segmentation, especially with respect to polyp segmentation and localization.

1. Introduction

Colorectal cancer (CRC) is the third cause of cancer death worldwide [1]. CRC arises from adenomatous polyps (adenomas) which are initially benign; however, over time, some of them can become malignant. Currently, the standard approach to reduce CRC-related mortality is to perform regular screening in search for polyps and colonoscopy is the screening tool of choice. During the examination, clinicians visually inspect the intestinal wall (see Figure 1(a) for an example of intestinal scene) in search of polyps. Once detected, they are resected and sent for histological analysis to determine their degree of malignancy and define the corresponding treatment the patient should undertake.

The main limitations of colonoscopy are its associated polyp miss rate (small/flat polyps or the ones hidden behind intestine folds can be missed [2]) and the fact that polyp's malignancy degree is only known after histological analysis. These drawbacks can be reduced by developing new colonoscopy modalities to improve visualization (e.g., high-definition imaging, narrow-band imaging (NBI) [3], and magnification endoscopes [4]) and/or by developing decision support systems (DSS) aiming to help clinicians in the different stages of the procedure. A clinically useful DSS should be able to detect, segment, and assess the malignancy degree (e.g., by optical biopsy [5]) of polyps during the colonoscopy procedure, following a similar pipeline to the one shown in Figure 1(b).

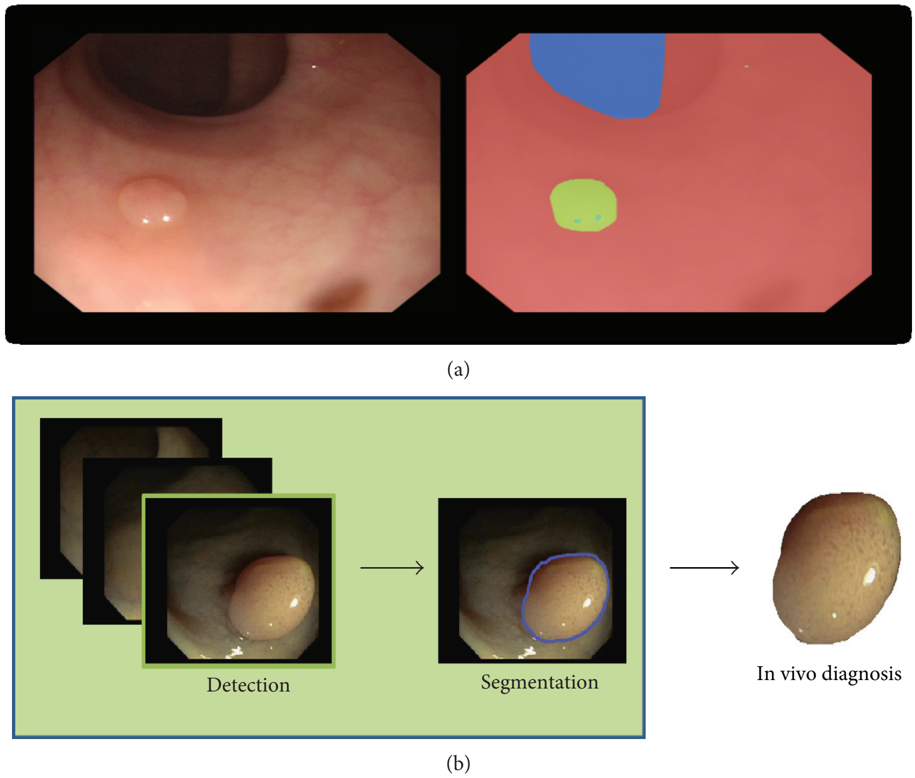


Figure 1: (a) Colonoscopy image and corresponding labeling: blue for lumen, red for background (mucosa wall), and green for polyp. (b) Proposed pipeline of a decision support system for colonoscopy.

The development of DSS for colonoscopy has been an active research topic during the last decades. The majority of available works on optical colonoscopy are focused on polyp detection (e.g., see [6–11]), and only few works address the problems of endoluminal scene segmentation.

Endoluminal scene segmentation is of crucial relevance for clinical applications [6, 12–14]. Polyp segmentation is important to define the area covered by a potential lesion that should be carefully inspected and possibly removed by clinicians. Moreover, having a system for accurate in vivo prediction of polyp histology might significantly improve clinical workflow. Lumen segmentation is relevant to help clinicians navigate through the colon during the procedure. Additionally, it can be used to establish quality metrics related to the degree of the colon wall that has been explored, since a weak exploration can lead to polyp overlooking. Finally, specular highlights have proven to be useful in reducing polyp detection false-positive ratio in the context of handcrafted methods [15].

In recent years, convolutional neural networks (CNNs) have become a de facto standard in computer vision, achieving state-of-the-art performance in tasks such as image classification, object detection, and semantic segmentation; and making traditional methods based on handcrafted features obsolete. Two major components in this groundbreaking progress were the availability of increased computational power (GPUs) and the introduction of large labeled datasets [16, 17]. Despite the additional difficulty of having limited amounts of labeled data, CNNs have successfully been applied to a variety of medical imaging tasks, by resorting to aggressive data augmentation techniques [18, 19]. More precisely, CNNs have excelled at semantic segmentation tasks in medical imaging, such as the EM ISBI 2012 dataset [20], BRATS [21], or MS lesions [22], where the top entries are built on CNNs [18, 19, 23–25]. Surprisingly, to the best of our knowledge, CNNs have not been applied to semantic segmentation of colonoscopy data. We associate this to the lack of large publicly available annotated databases, which are needed in order to train and validate such networks.

In this paper, we aim to overcome this limitation by introducing an extended benchmark of colonoscopy images created from the combination of the two largest public datasets of colonoscopy images [6, 26] and by incorporating additional annotations to segment lumen and specular highlights, with the hope of establishing a new strong benchmark for colonoscopy image analysis research. We provide new baselines on this dataset by training standard fully convolutional networks (FCNs) for semantic segmentation [27] and significantly outperforming, without any further postprocessing, prior results in endoluminal scene segmentation.

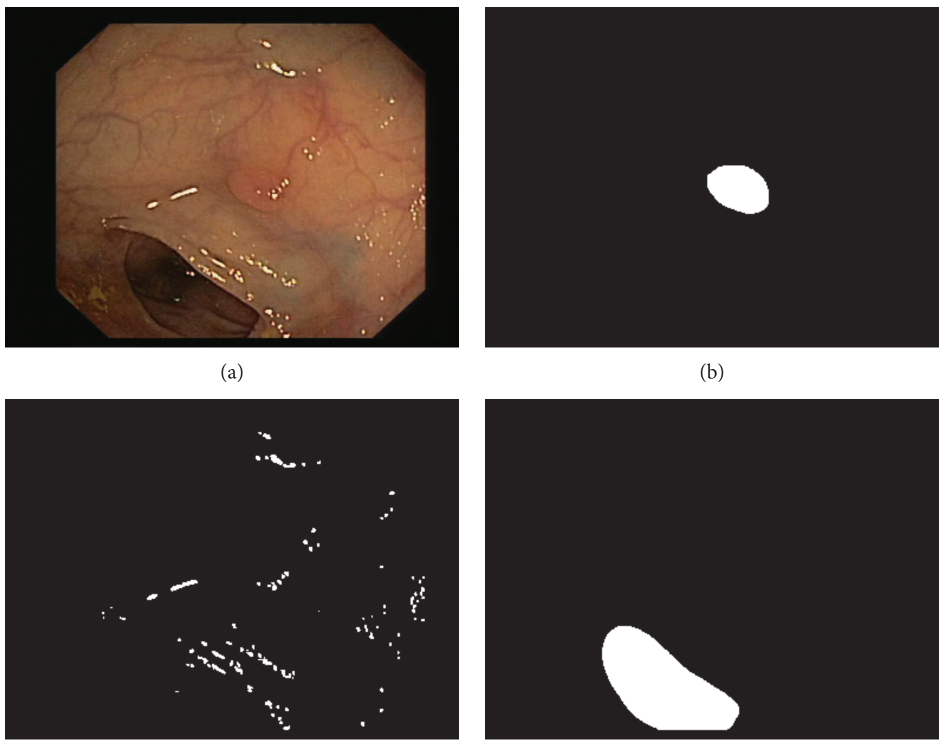
Therefore, the contributions of this paper are twofold:

(1) Extended benchmark for colonoscopy image segmentation

(2) New state-of-the-art in colonoscopy image segmentation.

Table 1: Summary of prior database content. All frames show at least one polyp.

Database	Number of patients	Number of seq.	Number of frames	Resolution	Annotations
CVC-ColonDB	13	13	300	500×574	Polyp, lumen
CVC-ClinicDB	23	31	612	384×288	Polyp
CVC-EndoSceneStill	36	44	912	500×574 & 384×288	Polyp, lumen, background, specularity, border (void)



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

Figure 2: Example of a colonoscopy image and its corresponding ground truth: (a) original image, (b) polyp mask, (c) specular highlights mask, and (d) lumen mask.

The rest of the paper is organized as follows. In Section 2, we present the new extended benchmark, including the introduction of datasets as well as the performance metrics. After that, in Section 3, we introduce the FCN architecture used as baseline for the new endoluminal scene segmentation benchmark. Then, in Section 4, we show qualitative and quantitative experimental results. Finally, Section 5 concludes the paper.

2. Endoluminal Scene Segmentation Benchmark

In this section, we describe the endoluminal scene segmentation benchmark, including evaluation metrics.

2.1. Dataset. Inspired by already published benchmarks for polyp detection, proposed within a challenge held in conjunction with MICCAI 2015 (http://endovis.grand-challenge.org) [28], we introduce a benchmark for endoluminal scene object segmentation.

We combine *CVC-ColonDB* and *CVC-ClinicDB* into a new dataset (*CVC-EndoSceneStill*) composed of 912 images obtained from 44 video sequences acquired from 36 patients.

(i) CVC-ColonDB contains 300 images with associated polyp masks obtained from 13 polyp video sequences acquired from 13 patients.

(ii) CVC-ClinicDB contains 612 images with associated polyp and background (here, mucosa and lumen) segmentation masks obtained from 31 polyp video sequences acquired from 23 patients.

We extend the old annotations to account for lumen, specular highlights with new hand-made pixel-wise annotations, and we define a void class for black borders present in each frame. In the new annotations, background only contains mucosa (intestinal wall). Please refer to Table 1 for dataset details and to Figure 2 for a dataset sample.

We split the resulting dataset into three sets: training, validation, and test containing 60%, 20%, and 20% images, respectively. We impose the constraint that one patient cannot be in different sets. As a result, the final training set contains 20 patients and 547 frames, the validation set contains 8 patients and 183 frames, and the test set contains 8 patients and 182 frames. The dataset is publicly available (http://www.cvc.uab.es/CVC-Colon/index.php/databases/cvc-endoscenestill/).

2.2. Metrics. We use Intersection over Union (IoU), also known as *Jaccard index*, and per pixel accuracy as segmentation metrics. These metrics are commonly used in medical image segmentation tasks [29, 30].

We compute the mean of per class IoU. Each per class IoU is computed over a validation/test set according to the following formula:

$$\mathrm{IoU}(\mathrm{PR}(\mathrm{class}), \mathrm{GT}(\mathrm{class})) = \frac{|\mathrm{PR}(\mathrm{class}) \cap \mathrm{GT}(\mathrm{class})|}{|\mathrm{PR}(\mathrm{class}) \cup \mathrm{GT}(\mathrm{class})|}, \quad (1)$$

where PR represents the binary mask produced by the segmentation method, GT represents the ground truth mask, $\cap$ represents set intersection, and $\cup$ represents set union.

We compute the mean global accuracy for each set as follows:

$$\mathrm{Acc}(\mathrm{PR}, \mathrm{GT}) = \frac{\#\mathrm{TP}}{\#\mathrm{pixels}}, \quad (2)$$

where TP represents the number of true positives.

Notably, this new benchmark might as well be used for the relevant task of polyp localization. In that case, we follow Pascal VOC challenge metrics [31] and determine that a polyp is localized if it has a high overlap degree with its associated ground truth, namely,

$$\mathrm{IoU}(\mathrm{PR}(\mathrm{polyp}), \mathrm{GT}(\mathrm{polyp})) > 0.5, \quad (3)$$

where the metric is computed for each polyp independently and averaged per set to give a final score.

3. Baseline

CNNs are a standard architecture used for tasks, where a single prediction per input is expected (e.g., image classification). Such architectures capture hierarchical representations of the input data by stacking blocks of convolutional, nonlinearity, and pooling layers on top of each other. Convolutional layers extract local features. Nonlinearity layers allow deep networks to learn nonlinear mappings of the input data. Pooling layers reduce the spatial resolution of the representation maps by aggregating local statistics.

FCNs [19, 27] were introduced in the computer vision and medical imaging communities in the context of semantic segmentation. FCNs naturally extend CNNs to tackle per pixel prediction problems, by adding upsampling layers to recover the spatial resolution of the input at the output layer. As a consequence, FCNs can process images of arbitrary size. In order to compensate for the resolution loss induced by pooling layers, FCNs introduce skip connections between their downsampling and upsampling paths. Skip connections help the upsampling path recover fine-grained information from the downsampling layers.

We implemented FCN8 architecture from [27] and trained the network by means of stochastic gradient descent with the rmsprop adaptive learning rate [32]. The validation split is used to early stop the training; we monitor mean IoU for validation set and use patience of 50. We used a minibatch size of 10 images. The input image is normalized in the range 0-1. We randomly crop the training images to 224×224 pixels. As regularization, we use dropout [33] of 0.5, as mentioned in the paper [27]. We do not use any weight decay.

As described in Section 2.1, colonoscopy images have a black border that we consider as a void class. Void classes do not influence the computation of the loss nor the metrics of any set, since the pixels marked as void class are ignored. As the number of pixels per class is unbalanced, in some experiments, we apply the median frequency balancing of [34].

During training, we experiment with data augmentation techniques such as random cropping, rotations, zooming, and sharing and elastic transformations.

4. Experimental Results

In this section, we report semantic segmentation and polyp localization results on the new benchmark.

4.1. Endoluminal Scene Semantic Segmentation. In this section, we first analyze the influence of different data augmentation techniques. Second, we evaluate the effect of having different numbers of endoluminal classes on polyp segmentation results. Finally, we compare our results with previously published methods.

4.1.1. Influence of Data Augmentation. Table 2 presents an analysis on the influence of different data augmentation techniques and their impact on the validation performance. We evaluate random zoom from 0.9 to 1.1, rotations from 0 to 180 degrees, shearing from 0 to 0.4, and warping with σ ranging from 0 to 10. Finally, we evaluate the combination of all the data augmentation techniques.

As shown in the table, polyps significantly benefit from all data augmentation methods, in particular, from warping. Note that warping applies small elastic deformation locally, accounting for many realistic variations in the polyp shape. Rotation and zoom also have a strong positive impact on the polyp segmentation performance. It goes without saying that such transformations are the least aggressive ones, since they do not alter the polyp appearance. Shearing is most likely the most aggressive transformation, since it changes the polyp appearance and might, in some cases, result in unrealistic deformations.

While for lumen it is difficult to draw any strong conclusions, it looks like zooming and warping slightly deteriorate the performance, whereas shearing and rotation slightly improve it. As for specular highlights, all

Table 2: FCN8 endoluminal scene semantic segmentation results for different data augmentation techniques. The results are reported on validation set.

Data augmentation	IoU background	IoU polyp	IoU lumen	IoU spec.	IoU mean	Acc mean
None	88.93	44.45	54.02	25.54	57.88	92.48
Zoom	89.89	52.73	51.15	37.10	57.72	90.72
Warp	90.00	54.00	49.69	**37.27**	58.97	90.93
Shear	89.60	46.61	54.27	36.86	56.83	90.49
Rotation	90.52	52.83	**56.39**	35.81	58.89	91.38
Combination	**92.62**	**54.82**	55.08	35.75	**59.57**	**93.02**

Table 3: FCN8 endoluminal scene semantic segmentation results for different numbers of classes. The results are reported on validation set. In all cases, we selected the model that provided best validation results (with or without class balancing).

Number of classes	IoU background	IoU polyp	IoU lumen	IoU spec.	IoU mean	Acc mean
4	92.07	39.37	**59.55**	**40.52**	57.88	92.48
3	92.19	50.70	56.48	—	66.46	92.82
2	**96.63**	**56.07**	—	—	**76.35**	**96.77**

Table 4: Results on the test set: FCN8 with respect to previously published methods.

	Data augmentation	IoU background	IoU polyp	IoU lumen	IoU spec.	IoU mean	Acc mean
FCN8 performance							
4 classes	None	86.36	38.51	**43.97**	32.98	50.46	87.40
3 classes	None	84.66	47.55	36.93	—	56.38	86.08
2 classes	None	**94.62**	50.85	—	—	**72.74**	**94.91**
4 classes	Combination	88.81	**51.60**	41.21	38.87	55.13	89.69
State-of-the-art methods							
[12, 13, 15]	—	73.93	22.13	23.82	**44.86**	41.19	75.58

the data augmentation techniques that we tested significantly boost the segmentation results. Finally, background (mucosa) shows only slight improvement when incorporating data augmentations. This is not surprising; given its predominance throughout the data, it could be even considered background.

Overall, combining all the discussed data augmentation techniques leads to better results in terms of mean IoU and mean global accuracy. More precisely, we increase the mean IoU by 4.51% and the global mean accuracy by 1.52%.

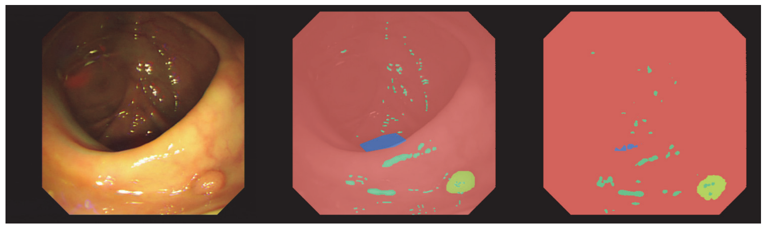
4.1.2. Influence of the Number of Classes. Table 3 presents endoluminal scene semantic segmentation results for different numbers of classes. As shown in the table, using more underrepresented classes such as lumen or specular highlights makes the optimization problem more difficult. As expected and contrary to handcrafted segmentation methods, when considering polyp segmentation, deep learning-based approaches do not suffer from specular highlights, showing the robustness of the learnt features towards saturation zones in colonoscopy images.

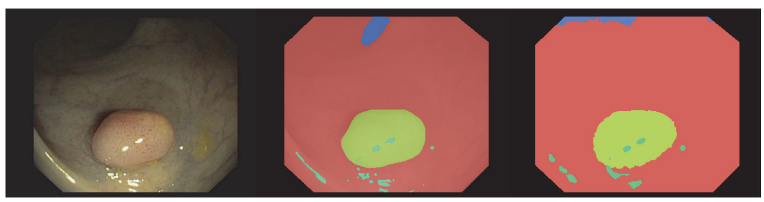
Best results for polyp segmentation are obtained in the 2-class scenario (polyp versus background). However, segmenting lumen is a relevant clinical problem as mentioned in Section 1. Results achieved in the 3-class scenario are very encouraging, with a IoU higher than 50% for both polyp and lumen classes.

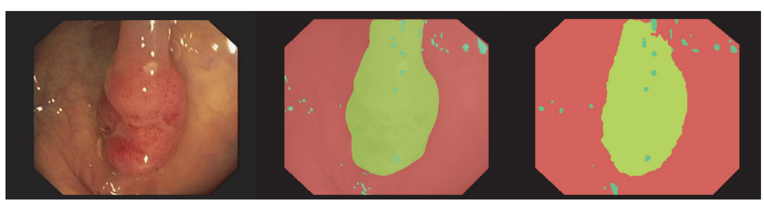
4.1.3. Comparison to State-of-the-Art. Finally, we evaluate the FCN model on the test set. We compare our results to the combination of previously published handcrafted methods: map-based method (1) for polyp segmentation and [12] a watershed-based method (2) for lumen segmentation and [15] (3) for specular highlights segmentation.

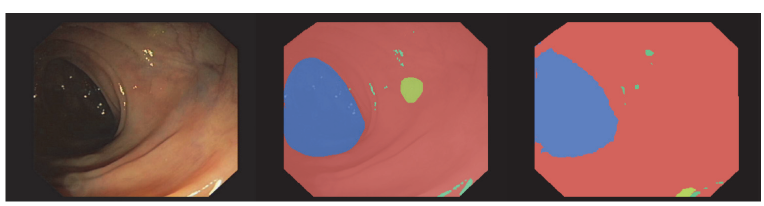
The segmentation results on the test set are reported in Table 4 and show a clear improvement of FCN8 over previously published methods. The following improvements can be observed when comparing previously published methods to the 4-class FCN8 model trained with data augmentation: 15% in IoU for background (mucosa), 29% in IoU for polyps, 18% in IoU for lumen, 14% in mean IoU, and 14% in mean accuracy. FCN8 is still outperformed by traditional methods when it comes to specular highlight class. However, it is important to note that specular highlight class is used by handcrafted methods to reduce false-positive ratio of polyp detection, and from our analysis, it looks like the FCN model is able to segment well polyps even when ignoring this class. For example, the best mean IoU of 72.74% and mean


(a)


(b)


(c)


(d)


(e)


(f)

Figure 3: Examples of predictions for 4-class FCN8 model. Each subfigure represents a single frame, a ground truth annotation, and a prediction image. We use the following color-coding in the annotations: red for background (mucosa), blue for lumen, yellow for polyp, and green for specularity. (a), (b), (c), (d) show correct polyp segmentation, whereas (e), (d) show incorrect polyp segmentation.

accuracy of 94.91% are obtained by the 2-class model without additional data augmentation.

Figure 3 shows qualitative results of the 4-class FCN8 model trained with data augmentation. From left to right, each row shows a colonoscopy frame, followed by the corresponding ground truth annotation and FCN8 prediction. Rows 1 to 4 show correct segmentation masks, with very clean polyp segmentation. Rows 5 and 6 show failure modes

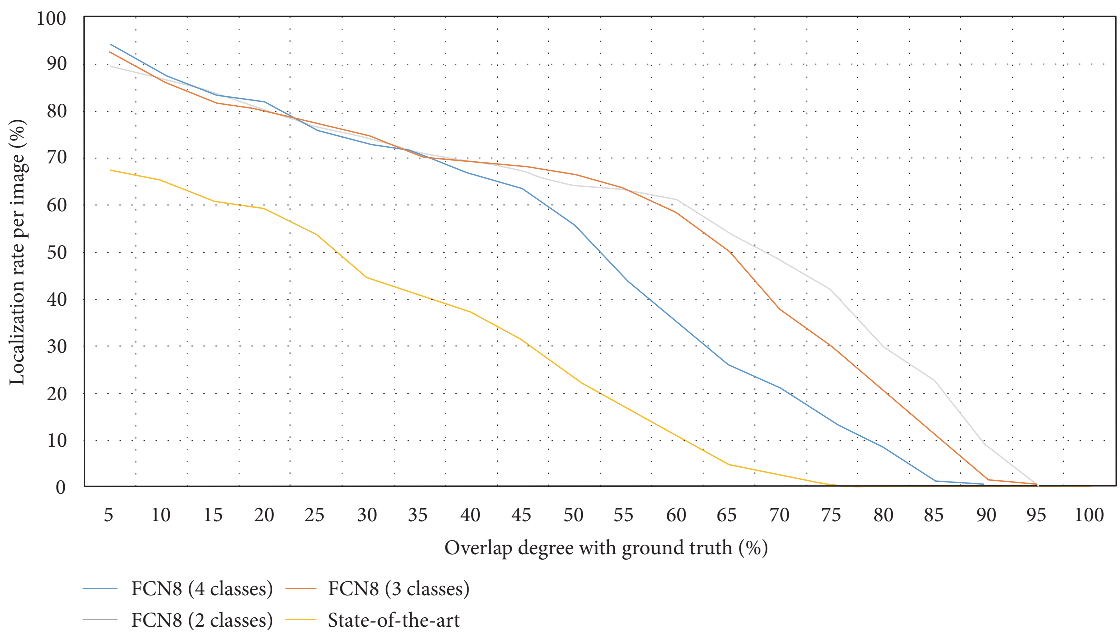


Figure 4: Localization rate of polyps as a function of IoU. The x-axis represents the degree of overlap between ground truth and model prediction. The y-axis represents the percentage of correctly localized polyps. Different color plots represent different models: FCN8 with 4 classes, FCN8 with 3 classes, and FCN8 with 2 classes and previously published method [13] (referred to as state-of-the-art in the plot).

Table 5: Summary of processing times achieved by the different methods studied in the paper. FCN results are the same for all four classes considered as segmentation of the four classes is done at the same time*.

Method	Polyp	Lumen	Specular highlights	Background
FCN	88 ms*	88 ms*	88 ms*	88 ms*
State-of-the-art	10000 ms	8000 ms	5000 ms	23000 ms

of the model, where polyps have been missed or undersegmented. In row 5, the small polyp is missed by our segmentation method while, in row 6, the polyp is undersegmented. All cases exhibit decent lumen segmentation and good background (mucosa) segmentation.

4.2. Polyp Localization. Endoluminal scene segmentation can be seen as a proxy to proper polyp detection in a colonoscopy video. In order to understand how well suited FCNs are to localize polyps, we perform a last experiment. In this experiment, we compute the polyp localization rate as a function of IoU between the model prediction and the ground truth. We can compute this IoU per frame, since our dataset contains a maximum of one polyp per image. This analysis describes the ability of a given method to cope with polyp appearance variability and stability on polyp localization.

The localization results are presented in Figure 4 and show a significant improvement when comparing FCN8 variants to the previously published method [13]. For example, when considering a correct polyp localization to have at least 50% IoU, we observe an increase of 40% in the polyp localization rate. As a general trend, we observe that architectures trained using a fewer number of classes achieve a higher IoU, though the polyp localization difference starts to be more visible when really high overlapping degrees are imposed. Finally, as one would expect, we observe that the architectures that show better results in polyp segmentation are the ones that show better results in polyp localization.

4.3. Towards Clinical Applicability. Sections 4.1.3 and 4.2 presented results of a comparative study between FCNs and previous state-of-the-art of endoluminal scene object segmentation in colonoscopy images. As mentioned in Section 1, we foresee several clinical applications, which can be built from the results of endoluminal scene segmentation. However, in order to be deployed in the exploration room, they must comply with real-time constraints apart from offering a good segmentation performance. In this case and considering videos recorded at 25 frames per second, a DSS should not take more than 40 ms to process an image in order not to delay the procedure.

Considering this, we have computed processing times for each of the approaches studied in this paper. Results are presented in Table 5.

As shown in the table, none of the presented approaches currently meet real-time constraints. Running the FCN8 inference on an NVIDIA Titan X GPU takes 88 ms per frame. Note that this could easily be addressed by taking advantage of recent research on model compression [35] by applying fancier FCN architectures that encourage feature reuse [36]. Alternatively, we could exploit the temporal component and build more sophisticated architectures that would take advantage of the similarities among consecutive frames.

Clearly, handcrafted methods take much longer to process one image. Moreover, they need to apply different methods to segment each class of interest, making them less clinically useful. Note that this is not the case for FCN-like architectures.

Despite computational constraints, FCNs' superior performance could lead to more reliable and impactful computer-assisted clinical applications, since they offer both a better performance and computational efficiency.

5. Conclusions

In this paper, we have introduced an extended benchmark for endoluminal scene semantic segmentation. The benchmark includes extended annotations of polyps, background (mucosa), lumen, and specular highlights. The dataset provides the standard training, validation, and test splits for machine learning practitioners and will be publicly available upon paper acceptance. Moreover, standard metrics for the comparison have been defined, with the hope to speed up the research in the endoluminal scene segmentation area.

Together with the dataset, we provided new baselines based on fully convolutional networks, which outperformed by a large margin previously published results, without any further postprocessing. We extended the proposed pipeline and used it as proxy to perform polyp detection. Due to the lack of nonpolyp frames in the dataset, we reformulated the task as polyp localization. Once again, we highlighted the superiority of deep learning-based models over traditional handcrafted approaches. As expected and contrary to handcrafted segmentation methods, when considering polyp segmentation, deep learning-based approaches do not suffer from specular highlights, showing the robustness of the learnt features towards saturation zones in colonoscopy images. Moreover, given that FCN not only excels in terms of performance but also allows for nearly real-time processing, it has a great potential to be included in future DSS for colonoscopy.

Knowing the potential of deep learning techniques, efforts in the medical imaging community should be devoted to gather larger labeled datasets as well as designing deep learning architectures that would be better suited to deal with colonoscopy data. This paper pretends to make a first step towards novel and more accurate DSS by making all code and data publicly available, paving the road for more researchers to contribute to the endoluminal scene segmentation domain.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the developers of Theano [37] and Keras [38]. The authors acknowledge the support of the following agencies for research funding and computing support: Imagia Inc.; Spanish government through funded Project AC/DC TRA2014-57088-C2-1-R and iVENDIS (DPI2015-65286-R); SGR Projects 2014-SGR-1506, 2014-SGR-1470, and 2014-SGR-135; CERCA Programme/Generalitat de Catalunya; and TECNIOspring-FP7-ACCI grant, FSEED, and NVIDIA Corporation for the generous support in the form of different GPU hardware units.

References

[1] Society, A.C, *Colorectal cancer*, 2016.

[2] A. Leufkens, M. Van Oijen, F. P. Vleggaar, and P. D. Siersema, "Factors inuencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.

[3] H. Machida, Y. Sano, Y. Hamamoto et al., "Narrow-band imaging in the diagnosis of colorectal mucosal lesions: a pilot study," *Endoscopy*, vol. 36, no. 12, pp. 1094–1098, 2004.

[4] M. Bruno, "Magnification endoscopy, high resolution endoscopy, and chromoscopy; towards a better optical diagnosis," *Gut*, vol. 52, Supplement 4, pp. iv7–iv11, 2003.

[5] H. K. Roy, M. J. Goldberg, S. Bajaj, and V. Backman, "Colonoscopic optical biopsy: bridging technological advances to clinical practice," *Gastroenterology*, vol. 140, no. 7, p. 1863, 2011.

[6] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[7] S. Gross, T. Stehle, A. Behrens et al., "A comparison of blood vessel features and local binary patterns for colorectal polyp classification," in *SPIE Medical Imaging*, article 72602Q, International Society for Optics and Photonics, 2009.

[8] S. Y. Park and D. Sargent, "Colonoscopic polyp detection using convolutional neural networks," in *SPIE Medical Imaging*, article 978528, International Society for Optics and Photonics, 2016.

[9] E. Ribeiro, A. Uhl, and M. Häfner, "Colonic polyp classification with convolutional neural networks," in *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 253–258, Dublin, 2016.

[10] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[11] N. Tajbakhsh, S. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2015.

[12] J. Bernal, D. Gil, C. Sánchez, and F. J. Sánchez, "Discarding non informative regions for efficient colonoscopy image analysis," in *International Workshop on Computer-Assisted and Robotic Endoscopy*, pp. 1–10, Springer, 2014.

[13] J. Bernal, J. M. Núñez, F. J. Sánchez, and F. Vilariño, "Polyp segmentation method in colonoscopy videos by means of MSA-DOVA energy maps calculation," in *MICCAI 2014 Workshop on Clinical Image-Based Procedures*, pp. 41–49, Springer, 2014.

[14] J. M. Núñez, J. Bernal, M. Ferrer, and F. Vilariño, "Impact of keypoint detection on graph-based characterization of blood vessels in colonoscopy videos," in *International Workshop on Computer-Assisted and Robotic Endoscopy*, pp. 22–33, Springer, 2014.

[15] J. Bernal, J. Sánchez, and F. Vilarino, "Impact of image preprocessing methods on polyp localization in colonoscopy frames," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7350–7354, Osaka, 2013.

[16] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," *CVPR09*, 2009.

[17] T. Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *European Conference on Computer Vision (ECCV)*, Zurich, 2014.

[18] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," 2016, http://arxiv.org/abs/1608.04117.

[19] 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, Springer, Cham, 2015.

[20] I. Arganda-Carreras, S. C. Turaga, D. R. Berger et al., "Crowdsourcing the creation of image segmentation algorithms for connectomics," *Frontiers in Neuroanatomy*, vol. 9, p. 142, 2015.

[21] B. Menze, A. Jakab, S. Bauer et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[22] M. Styner, J. Lee, B. Chin et al., "3d segmentation in the clinic: a grand challenge ii: Ms lesion segmentation," *Midas Journal*, vol. 2008, pp. 1–6, 2008.

[23] T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, "Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.

[24] H. Chen, X. Qi, J. Cheng, and P. A. Heng, "Deep contextual networks for neuronal structure segmentation," in *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, pp. 1167–1173, Phoenix, Arizona, USA, February 2016.

[25] M. Havaei, A. Davy, D. Warde-Farley et al., "Brain tumor segmentation with deep neural networks," 2015, http://arxiv.org/abs/1505.03540.

[26] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, 2015.

[28] J. Bernal, N. Tajbakhsh, F. J. Sanchez et al., "Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge," *IEEE Transactions on Medical Imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[29] K. H. Cha, L. Hadjiiski, R. K. Samala, H. P. Chan, E. M. Caoili, and R. H. Cohan, "Urinary bladder segmentation in ct urography using deep-learning convolutional neural network and level sets," *Medical Physics*, vol. 43, no. 4, pp. 1882–1896, 2016.

[30] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig, "A brain tumor segmentation framework based on outlier detection," *Medical Image Analysis*, vol. 8, no. 3, pp. 275–283, 2004.

[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[32] T. Tieleman and G. Hinton, "rmsprop adaptive learning," *COURSERA: Neural Networks for Machine Learning*, 2012.

[33] 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, no. 1, pp. 1929–1958, 2014.

[34] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," 2014, http://arxiv.org/abs/1411.4734.

[35] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: hints for thin deep nets," in *Proceedings of ICLR*, 2015.

[36] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," 2016, http://arxiv.org/abs/1608.06993.

[37] Theano Development Team, "Theano: a python framework for fast computation of mathematical expressions," 2016, http://arxiv.org/abs/1605.02688.

[38] F. Chollet, "Keras," 2015, https://github.com/fchollet/keras.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 3090343, 31 pages
https://doi.org/10.1155/2017/3090343

Review Article

A Review on Human Activity Recognition Using Vision-Based Method

Shugang Zhang,[1] Zhiqiang Wei,[1] Jie Nie,[2] Lei Huang,[1] Shuang Wang,[1] and Zhen Li[1]

[1]*College of Information Science and Engineering, Ocean University of China, Qingdao, China*
[2]*Department of Computer Science and Technology, Tsinghua University, Beijing, China*

Correspondence should be addressed to Zhen Li; lizhen0130@gmail.com

Received 22 February 2017; Accepted 11 June 2017; Published 20 July 2017

Academic Editor: Dong S. Park

Human activity recognition (HAR) aims to recognize activities from a series of observations on the actions of subjects and the environmental conditions. The vision-based HAR research is the basis of many applications including video surveillance, health care, and human-computer interaction (HCI). This review highlights the advances of state-of-the-art activity recognition approaches, especially for the activity representation and classification methods. For the representation methods, we sort out a chronological research trajectory from global representations to local representations, and recent depth-based representations. For the classification methods, we conform to the categorization of template-based methods, discriminative models, and generative models and review several prevalent methods. Next, representative and available datasets are introduced. Aiming to provide an overview of those methods and a convenient way of comparing them, we classify existing literatures with a detailed taxonomy including representation and classification methods, as well as the datasets they used. Finally, we investigate the directions for future research.

1. Introduction

Human activity recognition (HAR) is a widely studied computer vision problem. Applications of HAR include video surveillance, health care, and human-computer interaction. As the imaging technique advances and the camera device upgrades, novel approaches for HAR constantly emerge. This review aims to provide a comprehensive introduction to the video-based human activity recognition, giving an overview of various approaches as well as their evolutions by covering both the representative classical literatures and the state-of-the-art approaches.

Human activities have an inherent hierarchical structure that indicates the different levels of it, which can be considered as a three-level categorization. First, for the bottom level, there is an atomic element and these action primitives constitute more complex human activities. After the action primitive level, the action/activity comes as the second level. Finally, the complex interactions form the top level, which refers to the human activities that involve more than two persons and objects. In this paper, we follow this three-level categorization namely action primitives, actions/activities, and interactions. This three-level categorization varies a little from previous surveys [1–4] and maintains a consistent theme. Action primitives are those atomic actions at the limb level, such as "stretching the left arm," and "raising the right leg." Atomic actions are performed by a specific part of the human body, such as the hands, arms, or upper body part [4]. Actions and activities are used interchangeably in this review, referring to the whole-body movements composed of several action primitives in temporal sequential order and performed by a single person with no more person or additional objects. Specifically, we refer the terminology human activities as all movements of the three layers and the activities/actions as the middle level of human activities. Human activities like walking, running, and waving hands are categorized in the actions/activities level. Finally, similar to Aggarwal et al.'s review [2], interactions are human activities that involve two or more persons and objects. The additional person or object is an important characteristic of

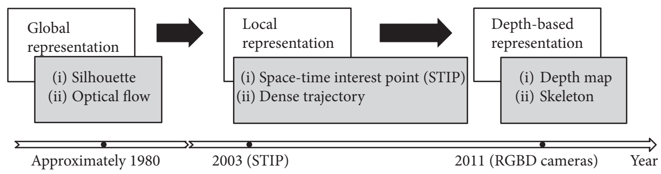


Figure 1: Research trajectory of activity representation approaches.

interaction. Typical examples of interactions are cooking which involves one person and various pots and pans and kissing that is performed by two persons.

This review highlights the advances of image representation approaches and classification methods in vision-based activity recognition. Generally, for representation approaches, related literatures follow a research trajectory of global representations, local representations, and recent depth-based representations (Figure 1). Earlier studies attempted to model the whole images or silhouettes and represent human activities in a global manner. The approach in [5] is an example of global representation in which space-time shapes are generated as the image descriptors. Then, the emergency of space-time interest points (STIPs) proposed in [6] triggered significant attention to a new local representation view that focuses on the informative interest points. Meanwhile, local descriptors such as histogram of oriented gradients (HOG) and histogram of optical flow (HOF) oriented from object recognition are widely used or extended to 3D in HAR area. With the upgrades of camera devices, especially the launch of RGBD cameras in the year 2010, depth image-based representations have been a new research topic and have drawn growing concern in recent years.

On the other hand, classification techniques keep developing in step with machine learning methods. In fact, lots of classification methods were not originally designed for HAR. For instance, dynamic time warping (DTW) and hidden Markov model (HMM) were first used in speech recognition [7, 8], while the recent deep learning method is first developed for large amount image classification [9]. To measure these approaches with same criterion, lots of activity datasets are collected, forming public and transparent benchmarks for comparing different approaches.

In addition to the activity classification approaches, another critical research area within the HAR scope, the human tracking approach, is also reviewed briefly in a separate section. It is widely concerned especially in video surveillance systems for suspicious behavior detection.

The writing of rest parts conforms to general HAR process flow. First, research emphases and challenges of this domain are briefly illustrated in Section 2. Then, effective features need to be designed for the representation of activity images or videos. Thus, Sections 3 and 4, respectively, review the global and local representations in conventional RGB videos. Depth image-based representations are discussed as a separate part in Section 5. Next, Section 6 describes the classification approaches. To measure and compare different approaches, benchmark datasets act an important role on which various approaches are evaluated. Section 7 collects recent human tracking methods of two dominant categories. In Section 8 we present representative datasets in different levels. Before we conclude this review and the future of HAR in Section 8, we classify existing literatures with a detailed taxonomy (Table 1) including representation and classification methods, as well as the used datasets aiming at a comprehensive and convenient overview for HAR researchers.

2. Challenges of the Domain

2.1. Intraclass Variation and Interclass Similarity. Different from speech recognition, there is no grammar and strict definition for human activities. This causes twofold confusions. On one hand, the same activity may vary from subject to subject, which leads to the intraclass variations. The performing speed and strength also increase the interclass gaps. On the other hand, different activities may express similar shapes (e.g., using a laptop and reading). This is termed as interclass similarity which is a common phenomenon in HAR. Accurate and distinctive features need to be designed and extracted from activity videos to deal with these problems.

2.2. Recognition under Real-World Settings

2.2.1. Complex and Various Backgrounds. While applications like video surveillance and fall detection system use static cameras, more scenarios adopt dynamic recording devices. Sports event broadcast is a typical case of dynamic recording. In fact, with the popularity of smart devices such as smart glasses and smartphones, people tend to record videos with embedded cameras from wearable devices anytime. Most of these real-world videos have complex dynamic backgrounds. First, those videos, as well as the broadcasts, are recorded in various and changing backgrounds. Second, realistic videos abound with occlusions, illumination variance, and viewpoint changes, which make it harder to recognize activities in such complex and various conditions.

2.2.2. Multisubject Interactions and Group Activities. Earlier research concentrated on low-level human activities such as jumping, running, and waving hands. One typical characteristic of these activities is having a single subject without any human-human or human-object interactions. However, in the real world, people tend to perform interactive activities with one or more persons and objects. An American football game is a good example of interaction and group activity where multiple players (i.e., human-human interaction) in a team protect the football (i.e., human-object interaction) jointly and compete with players in the other team. It is a challenging task to locate and track multiple subjects synchronously or recognize the whole human group activities as “playing football” instead of “running.”

Table 1: Taxonomy of activity recognition literatures.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Yamato et al. [94]	1992	Symbols converted from mesh feature vector and encoded by vector quantization (G)	HMM	RGB	Action/activity	Collected dataset: 3 subjects × 300 combinations	96% accuracy
Darrell and Pentland [92]	1993	View model sets (G)	Dynamic time warping	RGB	Action primitive	Collected instances of 4 gestures.	96% accuracy ("Hello" gesture)
Brand et al. [102]	1997	2D blob feature (G)	Coupled HMM (CHMM)	RGB	Action primitive	Collected dataset: 52 instances. 3 gestures × 17 times.	94.2% accuracy
Oliver et al. [97]	2000	2D blob feature (G)	(i) CHMM; (ii) HMM;	RGB	Interaction	Collected dataset: 11–75 training sequences +20 testing sequences. Organized as 5-level hierarchical interactions.	(i) 84.68 accuracy (average); (ii) 98.43 accuracy (average)
Bobick and Davis [17]	2001	Motion energy image & motion history image (G)	Template matching by measuring Mahalanobis distance	RGB	Action/activity	Collected dataset: 18 aerobic exercises × 7 views.	(a) 12/18 (single view); (b) 15/18 (multiple views)
Efros et al. [10]	2003	Optical flow (G)	K-nearest neighbor	RGB	Action/activity	(a) Ballet dataset; (b) tennis dataset; (c) football dataset	(a) 87.4% accuracy; (b) 64.3% accuracy; (c) 65.4% accuracy
Park and Aggarwal [103]	2004	Body model by combining an ellipse representation and a convex hull-based polygonal representation (G)	Dynamic Bayesian network	RGB	Interaction	Collected dataset: 56 instances. 9 interactions × 6 pairs of people.	78% accuracy
Schüldt et al. [105]	2004	Space-time interest points (L)	SVM	RGB	Action/activity	KTH dataset	71.7% accuracy
Blank et al. [5]	2005	Space-time shape (G)	Spectral clustering algorithm	RGB	Action/activity	Weizmann dataset	99.63% accuracy
Oikonomopoulos et al. [36]	2005	Spatiotemporal salient points (L)	RVM	RGB	Action/activity	Collected dataset: 152 instances. 19 activities × 4 subjects × 2 times.	77.63% recall
Dollar et al. [37]	2005	Space-time interest points (L)	(i) 1-nearest neighbor (1NN); (ii) SVM;	RGB	Action/activity	KTH dataset	(i) 78.5% accuracy (1NN); (ii) 81.17% accuracy (SVM)
Ke et al. [38]	2005	Integral videos (L)	Adaboost	RGB	Action/activity	KTH dataset	62.97% accuracy

Table 1: Continued.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Veeraraghavan et al. [93]	2005	Space-time shape (G)	Nonparametric methods by extending DTW	RGB	Action/activity	(a) USF dataset [154]; (b) CMU dataset [155]; (c) MOCAP dataset	No accuracy data presented.
Duong et al. [98]	2005	High level activities are represented as sequences of atomic activities; atomic activities are only represented using durations (–).	Switching hidden semi-Markov model (S-HSMM)	RGB	Interaction	Collected dataset: 80 video sequences. 6 high level activities.	97.5 accuracy (average accuracy; Coxian model)
Weinland et al. [20]	2006	Motion history volumes (G)	Principal component analysis (PCA) + Mahalanobis distance	RGB	Action/activity	IXMAS dataset [20]	93.33% accuracy
Lu et al. [49]	2006	PCA-HOG (L)	HMM	RGB	Action/activity	(a) Soccer sequences dataset [10]; (b) Hockey sequences dataset [156]	The implemented system can track subjects in videos and recognize their activities robustly. No accuracy data presented.
Ikizler and Duygulu [18]	2007	Histogram of oriented rectangles and encoded with BoVW (G)	(i) Frame by frame voting; (ii) global histogramming; (iii) SVM classification; (iv) dynamic time warping;	RGB	Action/activity	Weizmann dataset	100% accuracy (DTW)
Huang and Xu [19]	2007	Envelop shape acquired from silhouettes (G)	HMM	RGB	Action/activity; action primitive	Collected dataset: 9 activities × 7 subjects × 3 times × 3 views.	Subject dependent + view independent: 97.3% accuracy; subject independent + view independent: 95.0% accuracy; subject independent + view dependent: 94.4% accuracy
Scovanner et al. [46]	2007	3D SIFT (L)	SVM	RGB	Action/activity	Weizmann dataset	82.6% accuracy
Vail et al. [106]	2007	—	(i) HMM (ii) conditional random field	—	Interaction	Data from the hourglass and the unconstrained tag domains generated by robot simulator.	98.1% accuracy (CRF, hourglass); 98.5% accuracy (CRF, unconstrained tag domains)

Table 1: Continued.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Cherla et al. [21]	2008	Width feature of normalized silhouette box (G)	Dynamic time warping	RGB	Action/activity	IXMAS dataset [20]	80.05% accuracy; 76.28% accuracy (cross view)
Tran and Sorokin [25]	2008	Silhouette and optical flow (G)	(i) Naïve Bayes (NB); (ii) 1-nearest neighbor (1NN); (iii) 1-nearest neighbor with rejection (1NN-R); (iv) 1-nearest neighbor with metric learning (1NN-M)	RGB	Interaction; Action/activity	(a) Weizmann dataset; (b) UMD dataset [15]; (c) IXMAS dataset [20]; (d) collected dataset: 532 instances. 10 activities × 8 subjects.	(a) 100% accuracy; (b) 100% accuracy; (c) 81% accuracy; (d) 99.06% accuracy (1NN-M & L1SO)
Achard et al. [26]	2008	Semi-global features extracted from space-time micro volumes (L)	HMM	RGB	Action/activity	Collected dataset: 1614 instances. 8 activities × 7 subjects × 5 views.	87.39% accuracy (average)
Rodriguez et al. [91]	2008	Action MACH-maximum average correlation height (G)	Maximum average correlation height filter	RGB	Interaction; Action/activity	(a) KTH dataset; (b) collected feature films dataset: 92 kissing + 112 hitting/Slapping; (c) UCF dataset; (d) Weizmann dataset	(a) 80.9% accuracy; (b) 66.4% for kissing & 67.2% for hitting/slapping; (c) 69.2% accuracy; (d) reported a significant increase in algorithm efficiency, with no overall accuracy data presented
Kiaser et al. [30]	2008	Histograms of oriented 3D spatiotemporal gradients (L)	SVM	RGB	Interaction; Action/activity	(a) KTH dataset; (b) Weizmann dataset; (c) Hollywood dataset	(a) 91.4% (±0.4) accuracy; (b) 84.3% (±2.9) accuracy; (c) 24.7% precision
Willems et al. [39]	2008	Hessian-based STIP detector & SURF3D (L)	SVM	RGB	Action/activity	KTH dataset	84.26% accuracy
Laptev et al. [50]	2008	STIP with HOG, HOF are encoded with BoVW (L)	SVM	RGB	Interaction; Action/activity	(a) KTH dataset; (b) Hollywood dataset	(a) 91.8% accuracy; (b) 38.39% accuracy (average)
Natarajan and Nevatia [95]	2008	23 degrees body model (G)	Hierarchical variable transition HMM (HVT-HMM)	RGB	Action/activity; Action primitive	(a) Weizmann dataset; (b) gesture dataset in [157]	(a) 100% accuracy; (b) 90.6% accuracy
Natarajan and Nevatia [107]	2008	2-layer graphical model: top layer corresponds to actions in particular viewpoint; lower layer corresponds to individual poses (G)	Shape, flow, duration-conditional random field (SFD-CRF)	RGB	Action/activity	Collected dataset: 400 instances. 6 activities × 4 subjects × 16 views (×6 backgrounds).	78.9% accuracy

Table 1: Continued.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Ning et al. [108]	2008	Appearance and position context (APC) descriptor encoded by BoVW (L)	Latent pose conditional random fields (LPCRF)	RGB	Action/activity; Action primitive	HumanEva dataset	95.0% accuracy (LPCRF*init*)
Marszalek et al. [158]	2009	SIFT, HOG, HOF encoded by BoVW (L)	SVM	RGB	Interaction	Hollywood2 dataset	35.5% accuracy
Li et al. [76]	2010	Action graph of salient postures (D)	Non-Euclidean relational fuzzy (NERF) C-means & Hausdorf distance-based dissimilarity measure	Depth	Action/activity	MSR Action3D dataset	91.6% accuracy (train/test = 1/2); 94.2% accuracy (train/test = 2/1); 74.7% accuracy (train/test = 1/1 & cross subject)
Suk et al. [101]	2010	YIQ color model for skin pixels; histogram-based color model for face region; optical flow for tracking of hand motion (L)	Dynamic Bayesian network	RGB	Action primitive	Collected dataset: 498 instances. (a) 10 gestures × 7 subjects × 7 times (isolated gesture); (b) 8 longer videos contain 50 gestures (continuous gestures)	(a) 99.59% accuracy; (b) 84% recall & 80.77% precision
Baccouche et al. [124]	2010	SIFT descriptor encoded by BoVW (L)	Recurrent neural networks (RNN) with long short-term memory (LSTM)	RGB	Interaction	MICC-Soccer-Actions-4 dataset [159]	92% accuracy
Kumari and Mitra [29]	2011	Discrete Fourier transform on silhouettes (G)	K-nearest neighbor	RGB	Action/activity	(a) MuHaVi dataset; (b) DA-IICT dataset;	(a) 96% accuracy; (b) 82.6667% accuracy;
Wang et al. [51]	2011	Dense trajectory with HOG, HOF, MBH (L)	SVM	RGB	Interaction; Action/activity	(a) KTH dataset; (b) YouTube dataset; (c) Hollywood2 dataset; (d) UCF Sport dataset	(a) 94.2% accuracy; (b) 84.2% accuracy; (c) 58.3% accuracy; (d) 88.2% accuracy
Wang et al. [56]	2012	STIP with HOG, HOF are encoded with various encoding methods (L)	SVM	RGB	Interaction; Action/activity	(a) KTH dataset; (b) HMDB51 dataset	(a) 92.13% accuracy (Fisher vector); (b) 29.22% accuracy (Fisher vector)
Zhao et al. [77]	2012	Combined representations: (a) RGB: HOG & HOF upon space-time interest points (L) (b) depth: local depth pattern at each interest point (D)	SVM	RGB-D	Interaction	RGBD-HuDaAct dataset	89.1% accuracy

Table 1: Continued.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Yang et al. [78]	2012	DMM-HOG (D)	SVM	Depth	Action/activity	MSR Action3D dataset	95.83% accuracy (train/test = 1/2); 97.37% accuracy (train/test = 2/1); 91.63% accuracy (train/test = 1/1 & cross subject)
Xia et al. [84]	2012	Histograms of 3D joint locations (D)	HMM	Depth	Action/activity	(a) collected dataset: 6220 frames, 200 samples. 10 activities × 10 subjects × 2 times. (b) MSR Action3D dataset	(a) 90.92% accuracy; (b) 97.15% accuracy (highest); 78.97% accuracy (cross subject)
Yang and Tian [85]	2012	EigenJoints (D)	Naïve-Bayes-Nearest-Neighbor (NBNN)	Depth	Action/activity	MSR Action3D dataset	96.8% accuracy; 81.4% accuracy (cross subject)
Wang et al. [160]	2012	Local occupancy pattern for depth maps & Fourier temporal pyramid for temporal representation & actionlet ensemble model for characterizing activities (D)	SVM	Depth	Interaction; Action/activity	(a) MSR Action3D dataset; (b) MSR Action3DExt dataset; (c) CMU MOCAP dataset	(a) 88.2% accuracy; (b) 85.75% accuracy; (c) 98.13% accuracy
Wang et al. [53]	2013	Improved dense trajectory with HOG, HOF, MBH (L)	SVM	RGB	Interaction	(a) Hollywood2 dataset; (b) HMDB51 dataset; (c) Olympic Sports dataset [161]; (d) UCF50 dataset [162]	(a) 64.3% accuracy; (b) 57.2% accuracy; (c) 91.1% accuracy; (d) 91.2% accuracy
Oreifej and Liu [74]	2013	Histogram of oriented 4D surface normals (D)	SVM	Depth	Action/activity; Action primitive	(a) MSR Action3D dataset; (b) MSR Gesture3D dataset; (c) Collected 3D Action Pairs dataset	(a) 88.89% accuracy; (b) 92.45% accuracy; (c) 96.67% accuracy
Chaaraoui [88]	2013	Combined representations: (a) RGB: silhouette (G) (b) depth: skeleton joints (D)	Dynamic time warping	RGB-D	Action/activity	MSR Action3D dataset	91.80% accuracy
Ren et al. [152]	2013	Time-series curve of hand shape (G)	Dissimilarity measure based on Finger-Earth Mover's Distance (FEMD)	RGB	Action primitive	Collected dataset: 1000 instances. 10 gestures × 10 subjects × 10 times.	93.9% accuracy

Table 1: Continued.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Ni et al. [163]	2013	Depth-Layered Multi-Channel STIPs (L)	SVM	RGB-D	Interaction	RGBD-HuDaAct database	81.48% accuracy (codebook size = 512 & SPM kernel)
Grushin et al. [123]	2013	STIP with HOF (L)	Recurrent neural networks (RNN) with long short-term memory (LSTM)	RGB	Action/activity	KTH dataset	90.7% accuracy
Peng et al. [31]	2014	(i) STIP with HOG, HOF and encoded by various encoding methods; (L) (ii) iDT with HOG, HOF, MBHx, MBHy and encoded by various encoding methods (L)	SVM	RGB	Interaction	(a) HMDB51 dataset; (b) UCF50 dataset; (c) UCF101 dataset	Hybrid representation: (a) 61.1% accuracy; (b) 92.3% accuracy; (c) 87.9% accuracy
Peng et al. [32]	2014	Improved dense trajectory encoded with stacked Fisher kernal (L)	SVM	RGB	Interaction; Action/activity	(a) YouTube dataset; (b) HMDB51 dataset; (c) J-HMDB dataset	(a) 93.38% accuracy; (b) 66.79% accuracy; (c) 67.77% accuracy
Wang et al. [82]	2014	Local occupancy pattern for depth maps & Fourier temporal pyramid for temporal representation & actionlet ensemble model for characterizing activities (D)	SVM	Depth	Interaction; Action/activity	(a) MSR Action3D dataset; (b) MSR DailyActivity3D dataset; (c) Multiview 3D event dataset; (d) Cornell Activity Dataset [164]	(a) 88.2% accuracy; (b) 85.75% accuracy; (c) 88.34% accuracy (cross subject); 86.76% accuracy (cross view); (d) 97.06% (same person) 74.70% accuracy (cross person)
Simonyan and Zisserman [115]	2014	Spatial stream ConvNets & optical flow based temporal stream ConvNets (L)	SVM	RGB	Interaction	(a) HMDB51 dataset; (b) UCF101 dataset	(a) 59.4% accuracy; (b) 88.0% accuracy
Lan et al. [33]	2015	Improved dense trajectory with HOG, HOF, MBHx, MBHy enhanced with multiskip feature tracking (L)	SVM	RGB	Interaction	(a) HMDB51 dataset; (b) Hollywood2 dataset; (c) UCF101 dataset; (d) UCF50 dataset; (e) Olympic Sports dataset	(a) 65.1% accuracy (L = 3); (b) 68.0% accuracy (L = 3); (c) 89.1% accuracy (L = 3); (d) 94.4% accuracy (L = 3); (e) 91.4% accuracy (L = 3)
Shahroudy et al. [83]	2015	Combined representations: (a) RGB: dense trajectories with HOG, HOF, MBH (L) (b) Depth: skeleton joints (D)	SVM	RGB-D	Interaction	MSR DailyActivity3D	81.9% accuracy

Table 1: Continued.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Wang et al. [114]	2015	Weighted hierarchical depth motion maps (D)	Three-channel deep convolutional neural networks (3ConvNets)	Depth	Interaction; Action/activity	(a) MSR Action3D dataset; (b) MSR Action3DExt dataset; (c) UTKinect Action dataset [84]; (d) MSR DailyActivity 3D dataset; (e) Combined dataset of above	(a) 100% accuracy; (b) 100% accuracy; (c) 90.91% accuracy; (d) 85% accuracy; (e) 91.56% accuracy
Wang et al. [165]	2015	Pseudo-color images converted from DMMs (D)	Three-channel deep convolutional neural networks (3ConvNets)	Depth	Interaction; Action/activity	(a) MSR Action3D dataset; (b) MSR Action3DExt dataset; (c) UTKinect Action dataset [84]	(a) 100% accuracy; (b) 100% accuracy; (c) 90.91% accuracy
Wang et al. [117]	2015	Trajectory-pooled deep-convolutional descriptor and encoded by Fisher kernal (L)	SVM	RGB	Interaction	(a) HMDB51 dataset; (b) UCF101 dataset	(a) 65.9% accuracy; (b) 91.5% accuracy
Veeriah et al. [125]	2015	(i) HOG3D in KTH 2D action dataset; (L) (ii) skeleton-based features including skeleton positions, normalized pair-wise angels, offset of joint positions, histogram of the velocity, and pair-wise joint distances (D)	Differential recurrent neural network (dRNN)	RGBD	Action/activity	(a) KTH dataset; (b) MSR Action3D dataset	(a) 93.96% accuracy (KTH-1); 92.12% accuracy (KTH-2); (b) 92.03% accuracy
Du et al. [126]	2015	Representations of skeleton data extracted by subnets (D)	Hierarchical bidirectional recurrent neural network (HBRNN)	RGBD	Action/activity	(a) MSR Action3D dataset; (b) Berkeley MHAD Action dataset [166]; (c) HDM05 dataset [167]	(a) 94.49% accuracy; (b) 100% accuracy; (c) 96.92% (±0.50) accuracy
Zhen et al. [58]	2016	STIP with HOG3D and encoded with various encoding methods (L)	SVM	RGB	Interaction; Action/activity	(a) KTH dataset; (b) UCF YouTube dataset; (c) HMDB51 dataset	(a) 94.1% (Local NBNN); (b) 63.0% (improved Fisher kernal); (c) 30.5% (improved Fisher kernal)

Table 1: Continued.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Chen et al. [81]	2016	Action graph of skeleton-based features (D)	Maximum likelihood estimation	Depth	Action/activity	(a) MSR Action3D dataset; (b) UTKinect Action dataset	(a) 95.56% accuracy (cross subject); 96.1% accuracy (three subset evaluation); (b) 95.96% accuracy
Zhu et al. [87]	2016	Co-occurrence features of skeleton joints (D)	Recurrent neural networks (RNN) with long short-term memory (LSTM)	Depth	Interaction; Action/activity	(a) SBU Kinect interaction dataset [168]; (b) HDM05 dataset; (c) CMU dataset; (d) Berkeley MHAD Action dataset	(a) 90.41% accuracy; (b) 97.25% accuracy; (c) 81.04% accuracy; (d) 100% accuracy
Li et al. [116]	2016	VLAD for deep dynamics (G)	Deep convolutional neural networks (ConvNets)	RGB	Interaction; Action/activity	(a) UCF101 dataset; (b) Olympic Sports dataset; (c) THUMOS15 dataset [116]	(a) 84.65% accuracy; (b) 90.81% accuracy; (c) 78.15% accuracy
Berlin & John [119]	2016	Harris corner-based interest points and histogram-based features (L)	Deep neural networks (DNNs)	RGB	Interaction	UT Interaction dataset [169]	95% accuracy on set1; 88% accuracy on set2
Huang et al. [120]	2016	Lie group features (L)	Lie Group Network (LieNet)	Depth	Interaction; Action/activity	(a) G3D-Gaming dataset [170]; (b) HDM05 dataset; (c) NTU RGBD dataset [171]	(a) 89.10% accuracy; (b) 75.78% ± 2.26 accuracy; (c) 66.95% accuracy
Mo et al. [113]	2016	Automatically extracted features from skeletons data (D)	Convolutional neural networks (ConvNets) + multilayer perceptron	Depth	Interaction	CAD-60 dataset	81.8% accuracy
Shi et al. [55]	2016	Three stream sequential deep trajectory descriptor (L)	Recurrent neural networks (RNN) and deep convolutional neural networks (ConvNets)	RGB	Interaction; Action/activity	(a) KTH dataset; (b) HMDB51 dataset; (c) UCF 101 dataset [172]	(a) 96.8% accuracy; (b) 65.2% accuracy; (c) 92.2% accuracy
Yang et al. [79]	2017	Low-level polynormal assembled from local neighboring hypersurface normals and are then aggregated by Super Normal Vector (D)	Linear classifier	Depth	Interaction; Action/activity; Action primitive	(a) MSR Action3D dataset; (b) MSR Gesture3D dataset; (c) MSR Action Pairs3D dataset [173]; (d) MSR Daily Activity3D dataset	(a) 93.45% accuracy; (b) 94.74% accuracy; (c) 100% accuracy; (d) 86.25% accuracy

Table 1: Continued.

References	Year	Representation (global/local/depth)	Classification	Modality	Level	Dataset	Performance result
Jalal et al. [80]	2017	Multifeatures extracted from human body silhouettes and joints information (D)	HMM	Depth	Interaction; Action/activity	(a) Online self-annotated dataset [174]; (b) MSR Daily Activity3D dataset; (c) MSR Action3D dataset	(a) 71.6% accuracy; (a) 92.2% accuracy; (a) 93.1% accuracy

(a) (b)

Figure 2: Long-distance videos under real-world settings. (a) HAR in long-distance broadcasts. (b) Abnormal behaviors in surveillance.

2.2.3. Long-Distance and Low-Quality Videos. Long-distance and low-quality videos with severe occlusions exist in many scenarios of video surveillance. Large and crowded places like the metro and passenger terminal of the airport are representative occasions where occlusions happen frequently. Besides, surveillance cameras installed in high places cannot provide high-quality videos like present datasets in which the target person is clear and obvious. Though we do not expect to track everyone in these cases, some abnormal or crime-related behaviors should be recognized by the HAR system (Figure 2(b)). Another typical long-distance case is the football broadcast (Figure 2(a)). Due to the long distance of cameras, the subject is rather small which makes it difficult to analyze activities of the torso [10], and the relatively low quality of those long distance videos further increases the difficulty.

3. Global Representations

Global representations extract global descriptors directly from original videos or images and encode them as a whole feature. In this representation, the human subject is localized and isolated using background subtraction methods forming the silhouettes or shapes (i.e., region of interest (ROI)). Some global approaches encode ROI from which they derive corners, edges, or optical flow as descriptors. Other silhouette-based global representation methods stack the silhouette image along the time axis to form the 3D space-time volumes, then the volumes are utilized for representation. Besides, discrete Fourier transform (DFT) takes advantage of frequency domain information of ROI for recognition, also being a global approach. Global representation approaches were mostly proposed in earlier works and gradually outdated due to the sensitiveness to noise, occlusions, and viewpoint changes.

3.1. 2D Silhouettes and Shapes. To recognize the human activities in videos, an intuitive idea is to isolate the human body from the background. This procedure is called background subtraction or foreground extraction. The extracted foreground in the HAR is called silhouette, which is the region of interest and represented as a whole object in the global representation approach.

Calculating the background model is an important step before extracting silhouettes. Wren et al. [11] first proposed to model the background scene with Gaussian distribution. Koller et al. [12] pointed out that some foreground values update unduly and thus they introduced the selective background update strategy. Stauffer and Grimson [13] proposed to model the values of a particular background pixel as a mixture of Gaussians to replace the strategy of using only one Gaussian value in the previous approach. The Gaussian mixture model (GMM) has been applied widely but the introduction of expectation maximization (EM) algorithm increases the computational cost. To reduce the cost, k-means clustering algorithm is used to replace the EM algorithm with an insignificant loss of accuracy. It is worth mentioning that current RGBD cameras make it easy to obtain the silhouette by using the depth data provided by depth sensors.

Besides the silhouette representation, the 2D shape of the silhouette can be used as a feature as well. Veeraraghavan et al. [14] emphasized the effectiveness of shape features. In their experiments, shape and kinematics that are being considered as two important cues in human motion were evaluated. Tests on both the gait-based human identification and the activity recognition indicate that shape plays a more important role. Veeraraghavan et al. then used this shape representation in their following work [15].

Bobick and Davis [16, 17] stacked the silhouettes as two components for recognizing activities, respectively, the motion-energy image (MEI) and the motion-history image (MHI), which are both 2D representations.

In [18], oriented rectangular patches are extracted over the silhouettes. Spatial oriented histograms are then formed to represent the distribution of these rectangular patches. Those descriptors are finally used to recognize activities.

Extracting silhouettes from a single view is hard to satisfy view invariant property. To alleviate the influence of viewpoint changes, multiple cameras can be used to extract silhouettes in different viewpoints. Xu and Huang [19] proposed an “envelop shape” representation using two orthogonally placed cameras, which is robust to view changes of yaw rotation. Weinland et al. [20] made the same assumption that only the variations in viewpoints around the central vertical axis of the human body need to be considered. Motion history volumes (MHVs) were derived by stacking 4D silhouettes from four orthogonal cameras. In [21], a data fusion method was proposed, calculating the minimum

DTW score between the test template and the two orthogonal view training templates.

3.2. Optical Flow. Optical flow is an effective way to extract and describe silhouettes for a dynamic background. Lucas-Kanade-Tomasi (LKT) feature tracker [22, 23] can be used to obtain the optical flow. Lu et al. [24] used a LKT feature tracker approach to track joints in key frames and actual frames. Each activity is represented as a posture sequence, and each key posture is recorded in a key frame. Specific posture in actual frames can be recognized by finding correspondence between the actual and key frame. The recognized posture from the actual frame is compared to the key posture frame by mapping body locations, and the matched posture sequences are confirmed as the activity.

For recognizing human activities at a distance (i.e., the football broadcast video), Efros et al. [10] introduced a descriptor based on computing the optical flow to describe the "small" football players in person-centered images. Obviously, the background is dynamic due to the movement of players which makes it hard to model for background subtraction.

Tran and Sorokin [25] combined silhouettes and optical flow features together. Normalized bounding box is scaled to capture the region of the human body, and the optical flow measurements within the box are split into horizontal and vertical channels, while the silhouette gives the third channel. Subwindows are further divided to calculate histograms, and concatenating histograms of all 3 channels form the final descriptor.

3.3. 3D Space-Time Volumes (STVs). An activity video can be seen as a series of images that contain activity sequences. Concatenating all frames along the time axis forms the 3D space-time volume (STV) which has three dimensions including two spatial dimensions X and Y and one temporal dimension T. Representations based on STVs expect to capture the additional dynamic information which the spatial representation methods cannot obtain due to the absence of time dimension. Constructing STVs for different activities is a global representation method. However, the STV sometimes combines with local features to build the final feature sets.

Blank et al. [5] first introduced the space-time shape to represent human activities. Space-time shape is obtained by only stacking the silhouette regions within images. However, due to the nonrigidity of the constructed 3D space-time shapes and inherent difference between space and time dimensions, traditional 3D shape analysis cannot be applied to the space-time activity shapes. Thus, the solution of the Poisson equation is used to derive local space-time saliency and orientation features.

Achard et al. [26] generated semiglobal features named space-time micro volumes from image sequence to deal with performances of different temporal durations. Motivated by seeking the common underlying induced motion fields of sequences of the same behaviors, Shechtman et al. [27] proposed an approach to compare volumes according to their patches. This method requires no prior modeling or learning of activities, being able to handle the complex dynamic scenes and detect multiple activities that occur simultaneously within the camera view. Their method is partially invariant to the changes in scale and orientation.

In [28], the input videos are segmented into space-time volumes using mean shift clustering technique. These over-segmented regions, which are termed "super-voxels," are then matched using a proposed shape-matching technique, which is compared to the traditional silhouette matching methods. Unlike the previous silhouette-based approaches, the proposed shape-based representation does not require background subtraction nor explicit background models. To avoid the shortages of the shape-matching methods that are ignoring features inside the shape, Shechtman and Irani's flow-based features [27] are further incorporated.

3.4. Discrete Fourier Transform (DFT). The DFT of image frame is another global feature that contains the intensity information of the foreground object (i.e., the region of the subject's body) provided that the foreground object intensity is different from the background. Kumari and Mitra [29] took advantage of this hypothesis and proposed a DFT-based approach, obtaining information about the geometric structure of the spatial domain foreground object. Normalized image frame is divided into small size blocks within which the average of all the DFT values is calculated. Finally the K-nearest neighbor (KNN) is applied to classify the DFT features and generate the activity classification result. The extracted DFT feature is novel compared to the previous work; however, its performance is restricted to simple backgrounds. The background in their test video datasets is almost blank.

4. Local Representations

Instead of extracting the silhouette or STV and encoding them as a whole, local representations process activity video as a collection of local descriptors. They focus on specific local patches which are determined by interest point detectors or densely sampling [30]. Most existing local features are proved to be robust against noise and partial occlusions comparing to global features. Local features are then normally combined with the bag-of-visual-words (BoVW) model and yield the general pipeline of current state-of-the-art local representation approaches [31]. Oriented from bag-of-words (BoW), BoVW-based local representation mainly contains four steps: feature extraction, codebook generation, feature encoding, and pooling and normalization. We follow [32] and state a traditional BoVW pipeline here: interest points and local patches are first obtained by detectors or densely sampled. Then local features are extracted from those interest points or patches. Next, a visual dictionary (i.e., codebook) is learned in training set by k-means or Gaussian mixture model (GMM), the original high-dimension descriptors are clustered, and the center of each cluster is regarded as a visual codeword. After that, local features are encoded and pooled. Finally, the pooled vectors are normalized as video representation. Among these steps, the development of more elaborately designed low-level features and more sophisticated encoding methods are the two chief reasons for the great achievements in this field [32, 33], so in this part, we review the feature extraction methods in Section 4.1 and Section 4.2, as well as the encoding methods in Section 4.3.

4.1. Spatiotemporal Interest Point Detector. An intuitive thought of local representation is to identify those interest points that contain high information contents in images or videos. Harris and Stephens [34] first proposed effective 2D interest point detectors, the well-known Harris corner detector, which is extensively used in object detection. Then, Laptev and Lindeberg [6] proposed the 3D space-time interest points (STIPs) by extending Harris detectors. Spatial interest points in images are extended to spatiotemporal local structures in videos where the image values have significant local variations in both space and time. The spatiotemporal extents of the detected points are estimated by maximizing a normalized spatiotemporal Laplacian operator over spatial and temporal scales.

Saliency can also be used to detect interest points. Saliency means that certain parts of an image are preattentively distinctive and are immediately perceivable [35]. The spatiotemporal salient point can be regarded as an instance of the spatiotemporal interest point since both of them are informative and contain significant variations. The 2D salient point detection was first proposed by Kadir and Brady in [35]. Oikonomopoulos et al. [36] extended the 2D saliency to 3D spatiotemporal salient points that are salient both in space and time field. The salient points are successfully used as local features in their proposed activity classification scheme. Blank et al. [5] used the solution to Poisson equation to extract local space-time saliency of moving parts in the space-time shape. The detected salient points along with the local orientation and aspect ratios of shapes are calculated as local features.

Although these methods achieved remarkable results in HAR, one common deficiency is the inadequate number of stable interest points. In fact, the trade-off between the stability of those points and the number of points found is difficult to control. On one hand, the “right” and “discriminative” (i.e., stable) interest points are rare and difficult to be identified. As stated in [37], the direct 3D counterparts to commonly used 2D interest point detectors are inadequate, and true spatiotemporal corners are quite rare in certain applications. On the other hand, false alarms occur frequently due to various factors such as unintentional appearance changes. Ke et al. [38] illustrated two instances to point out that original detectors may fail in situations where the motions contain no sharp extrema; however, these detectors can be triggered falsely by the appearance of shadows and highlights in video sequences.

Besides the inherent properties of sparse interest points, many of the mentioned methods are inefficient. Therefore, these methods are restricted to the detection of a small number of points, or limited to low-resolution videos [39]. Here, we introduce some works either efficiency-enhanced or increasing number of stable interest points in response to the mentioned deficiency.

Dollar et al. [37] observed the rarity of the spatiotemporal interest points and the consequent problems of it in the recognition scheme. To find more 3D interest points in cuboids of space and time for activity recognition, the response function calculated by the separable linear filters is applied. The filtering is applied separately on the spatial and temporal dimensions, that is, 2D Gaussian smoothing kernel applied in spatial dimensions, and 1D Gabor filters applied in temporal dimension. Number of interest points increases using their detectors. Ke et al. [38] doubted the assumption that one can reliably detect a sufficient number of stable interest points in the video sequence. They extended the notion of rectangle features [40] into spatiotemporal volumetric features and applied the proposed framework on the video's optical flow. Their classifier is not limited to the sparseness nor affected by the instability of detected points.

Aiming at detecting interest points in an efficient way, Willems et al. [39] presented a dense, scale-invariant yet efficient spatiotemporal interest point detector with minimal effect on the computation time. First, point localization and scale selection are combined in a direct way using the determinant of the 3D Hessian matrix, therefore removing the time-consuming iterative scheme [41]. Further, building on Ke et al.'s work [38], an implementation scheme using integral video is developed to compute scale-invariant spatiotemporal features efficiently. Using a completely different idea, Oshin et al. [42] proposed to learn a classifier capable of detecting interest points in a novel video, given examples of the type of interest point that wish to get within a training video. The spatiotemporal Fern classifier (i.e., a seminaïve Bayesian classifier in [43]) is trained to recognize spatiotemporal interest points and thus achieves a high efficiency in constant time regardless of original detector complexity.

4.2. Local Descriptors. Local descriptors are designed to describe the patches that sampled either densely or at the interest points [1]. Effective descriptors are considered to be discriminative for the target human activity events in videos and robust to occlusion, rotation, and background noise.

Laptev [41] represented their 3D Harris corner by computing local, spatiotemporal N-jets as the descriptor. The descriptor is scale-invariant since they estimate the spatiotemporal extents of detected events by maximizing a normalized spatiotemporal Laplacian operator over spatial and temporal scales. Moreover, the proposed descriptors are proved to be robust to occlusions and dynamic cluttered backgrounds in the human motion analysis.

Similar to works of extending 2D interest point detector into spatiotemporal domain, such as the Harris corner detector [34] and the extended spatiotemporal one [41], many spatiotemporal descriptors were proposed by extending mutual image descriptors as well. We briefly review these works including both the original spatial descriptors and the spatiotemporal version of them.

Lowe proposed the scale-invariant feature transform (SIFT) in 1999 [44] and further improved it in 2004 [45]. It is widely used in local representation due to its scale and rotation invariance, as well as the robustness to affine distortion, changes in 3D viewpoint, addition of noise, and change in illumination. Scovanner et al. [46] introduced a 3D SIFT descriptor and used it in HAR. The 2D gradient magnitude and orientation are extended in 3D formulation; thus, creating the subhistograms encode the 3D SIFT descriptor. The videos are then described as a bag of spatiotemporal words using the 3D SIFT descriptor. Moreover, a feature grouping histogram which groups the co-occurred words out of the original one is used to build a more discriminative action video representation and finally used for classification.

The speeded-up robust features (SURF) [47] approach is a scale and rotation invariant detector and descriptor. The most important property of SURF is the improvement of efficiency comparing to previous approach. In the interest point detection, the approach applies the strategy that analyzing the input image at different scales to guarantee invariance to scale changes. Taking computation time into account, a very basic Hessian-matrix approximation which lends itself to the use of integral images is used for interest point detection, and it reduced the computation time dramatically. Next, a rotation and scale-invariant descriptor is provided for the detected interest point. The SURF approach builds on the distribution of first-order Haar-wavelet responses within the interest point neighborhood, in contrast with SIFT that extracts gradient information. Furthermore, integral images are exploited for speed. The introduction of indexing step based on the sign of the Laplacian further increases the robustness of descriptor and the matching speed.

An extended 3D SURF descriptor was implemented by Willems et al. [39]. Both of the 2D and 3D SURF used Haar-wavelet responses; however, the 3D SURF store the vector of the 3 axis responses instead of including the sums over the absolute values since the latter proved to be of no significant benefit but doubling the descriptor size.

Dalal and Triggs [48] proposed the histogram of oriented gradients (HOG) descriptor and achieved great success in human detection with linear SVM classifier. The good performance is due to the fact that the HOG's density distribution of local intensity gradients or edge directions can well characterize the local object appearance and shape of target objects.

Lu and Little et al. [49] presented the PCA-HOG descriptor which projects the original histogram of oriented gradients (HOG) descriptor to a linear subspace by principle component analysis (PCA). The descriptor was used to represent athletes to solve the problem of tracking and activity recognition simultaneously. Using HOG and HOF (histogram of flow) descriptor, Laptev et al. [50] completes a similar but more challenging activity recognition task as those activities are extracted from movies.

Klaser et al. [30] generalized the HOG descriptor to video sequences and proposed the HOG3D. Integral images are extended to integral videos for efficient 3D gradient computation. Polyhedrons are utilized for orientation quantization as an analogy of polygons in 2D space HOG. Optimized parameters for activity recognition have also been explored in their work.

Early spatiotemporal methods adopt a perspective of regarding the video as x-y-t 3D volumes [30, 39, 46]. However, recent feature trajectory approach considers the spatial dimensions x-y very different from the temporal dimension t. This approach detects the x-y interest points from video frames and then tracking them through video sequences as a trajectory. For detecting interest point, classic 2D detectors such as HOG and HOF are still used. In this review, we treat the feature trajectory as a special kind of the spatiotemporal descriptors where the time dimension is used to concatenate those 2D interest points.

Wang et al. [51] proposed dense trajectories by densely sampling points. Avoiding extracting points frame by frame and concatenating them, Wang et al. firstly extracted dense optical flow using Farneback's algorithm [52], then points can be densely tracked along the trajectory without additional cost. HOG and HOF are computed along the dense trajectories as the descriptors. Dense trajectories were further improved in [53]. The camera motion, as a main obstacle for extracting target trajectories from humans or objects of interests, was highlighted and was tried to be removed. The authors first match feature points using two complementary descriptors (i.e., SURF and dense optical flow), then estimate the homography using RANSC [54]. Through this approach, the camera motion is explicitly identified and removed. However, in some cases where humans dominate the frame, the target human motion may also generate inconsistent camera motion match. To solve this problem, a human detector is further explored to remove the inconsistent matches within the detected human areas. Improved descriptors achieved significant performance on challenge datasets, such as Hollywood2 where camera motions were used abundantly. Shi et al. [55] presented a sequential deep trajectory descriptor (sDTD) on the dense trajectory basis to capture the long-term motion information. The dense trajectories are projected into two-dimensional planes and a CNN-RNN network is employed to learn an effective representation for long-term motion.

4.3. Feature Encoding Methods. The STIP-based descriptors or other elaborately designed descriptors are all referred as local features. Local features are then encoded with feature encoding methods to represent activities and the encoded features are subsequently fed into pretrained classifiers (e.g., SVM) [32]. Encoding feature is a key step for constructing BoVW representation and utilizing an appropriate encoding method can significantly improve the recognition accuracy [56]. Here, we summarize the common feature encoding methods in recent literatures in Table 2. The number of citations for each description paper is also provided to facilitate measurement of their influences.

Several evaluations [56–58] have been conducted to compare the performance of recent encoding methods. Chatfield et al. [57] compared five encoding methods including LLC, SVC, FV, KCB, and the standard spatial histograms baseline. Experiments over PASCAL VOC 2007 and Caltech 101 show that FV performs best. Wang et al. [56] drew the same conclusion on KTH dataset and HMDB51 dataset. Also, a most recent evaluation [58] showed a consistent finding on UCF-YouTube and HMDB51 datasets, though slightly slower than local NBNN on KTH.

Table 2: Feature encoding methods.

Method	Proposed	Description paper, the number of citations
Vector quantization (VQ)/hard assignment (HA)	Sivic et al. (2003)	[59], 5487
Kernal codebook coding (KCB)/soft assignment (SA)	Gemert et al. (2008)	[60], 586; [61], 761
Spase coding (SPC)	Yang et al. (2009)	[62], 2529
Local coordinate coding (LCC)	Yu et al. (2009)	[63], 614
Locality-constrained linear coding (LLC)	Wang et al. (2010)	[64], 2410
Improved Fisher kernel (iFK)/Fisher vector (FV)	Perronnin et al. (2010)	[65], 1590
Triangle assignment coding (TAC)	Coates et al. (2010)	[66], 976
Vector of locally aggregated descriptors (VLAD)	Jegou et al. (2010)	[67], 1135; [68], 710;
Super vector coding (SVC)	Zhou et al. (2010)	[69], 459
Local tangent-based coding (LTC)	Yu et al. (2010)	[70], 122
Localized soft assignment coding (LSC/SA-*k*)	Liu et al. (2011)	[71], 398
Salient coding (SC)	Huang et al. (2011)	[72], 131
Group salient coding (GSC)	Wu et al. (2012)	[73], 33
Stacked Fisher vectors (SFV)	Peng et al. (2014)	[32], 149

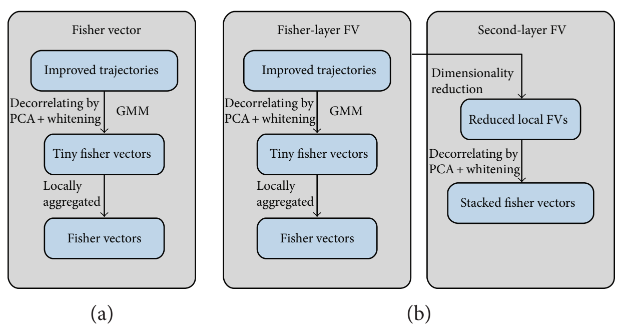


Figure 3: Pipeline of Fisher vector and Stacked fisher vector. (a) Fisher vector. (b) Stacked fisher vector.

Further exploration has been conducted to match the best local feature with FV. In [31], six representative methods including VQ, SA-*k*, LLC, FV, VLAD, and SVC are evaluated for two widely used local features, STIPs and improved dense trajectories (iDTs). The experiment results demonstrate that the iDT together with the FV yields the best performance on the test datasets. Wang et al. who proposed the iDT also verified the best performance of iDT and FV in their work [53].

Recent stacked Fisher vectors [32] further improved the performance of iDT + FV and achieved superior performance when combining traditional FV. Evaluation on the YouTube, J-HMDB, and HMDB51 datasets demonstrates that it has become the state-of-the-art method. Pipelines of SFV and corresponding FV are given in Figure 3.

The core idea of both FV and SFV is trying to catch more statistical information from images; in contrast, BoVW only retains the zero order statistics. Take an l-dimension local descriptor as an example. Assuming that the size of prelearned GMM is K (K is the size of codebook). For the conventional BoVW, the final encoded feature is K-dimension histograms that indicate the frequency of codewords. However, FV can obtain a $2Kd$-dimension (d is the Gaussian distribution dimension). In another word, FV retained more information (i.e., high-order statistics) regarding to same size of codebooks.

SFV further improved FV owing to a simple and intuitive reason that SFV densely calculated local features by dividing and scanning multiscale subvolumes. The main challenge is the holistic combination of those local FVs since encoding them using another FV directly is impossible because of the high dimension of them ($2Kd$-dimension). Thus, a max-margin method is tactfully used to reduce dimensionality. As the local FVs are more densely sampled than the conventional FV and consequently contain more high order statistics, therefore, iDT with SFV achieves even better result than the state-of-the-art iDT with FV.

5. Depth-Based Representations

Previous research of HAR mainly concentrates on the video sequences captured by traditional RGB cameras. Depth cameras, however, have been limited due to their high cost and complexity of operation [74]. Thanks to the development of low-cost depth sensors such as Microsoft Kinect [75], an

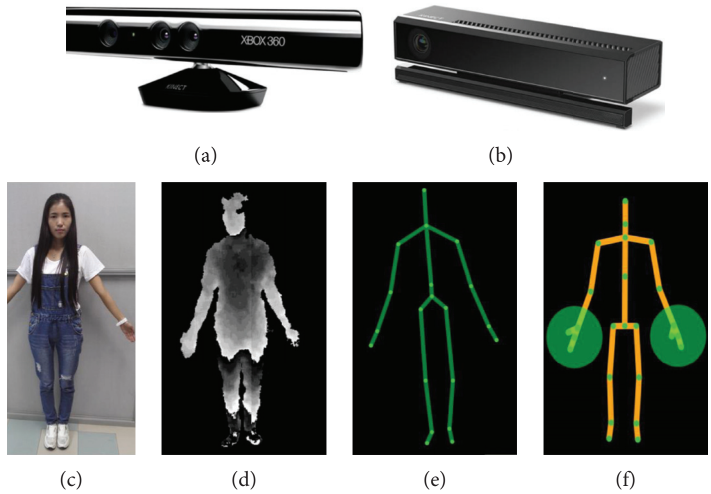


Figure 4: Kinect RGBD cameras and their color images, depth maps, skeletal information. (a) Kinect v1 (2011). (b) Kinect v2 (2014). (c) Color image. (d) Depth map. (e) Skeleton captured by Kinect v1. (f) Skeleton captured by Kinect v2.

affordable and easier way to access the depth maps is provided. Furthermore, Kinect SDK released the application that can directly obtain the skeletal joint positions in real-time (adopting algorithms in [75]). The available depth maps and the skeletal information (see Figure 4) vigorously contributed to the computer vision community. These two features and their derivative features also triggered a wide interest to solve HAR problems using depth-based solutions, replacing conventional RGB-based methods, or acting as supplements to enhance the RGB-based methods. In this section, we separately reviewed the recent advance of activity representations using depth maps or skeletons.

5.1. Representations Based on Depth Maps. Depth maps contain additional depth coordinates comparing to conventional color images and are more informative. Approaches presented in this section regard depth maps as spatiotemporal signals and extract features directly from them. These features are either used independently or combined with RGB channel to form multimodal features.

Li et al. [76] employed the action graph model, which represents activities using several salient postures serving as nodes in action graph. All activities share same posture sets and each posture is characterized as a bag of 3D points from the depth maps. However, involving all the 3D points is computationally expensive; thus, a simple and effective method to sample the representative 3D points is proposed, achieving over 90% recognition accuracy by sampling approximately 1% points according to their report.

Zhao et al. [77] proposed a framework of combing RGB and depth map features for HAR and presented an optimal scheme. For the RGB channels, spatiotemporal interest points are generated solely from it and the HOG and HOF are calculated to form the RGB based descriptors. For the depth channel, they proposed a depth map-based descriptor called local depth pattern (LDP), which simply calculates the difference of average depth values between a pair of cells within the STIP surrounding region.

Yang et al. [78] proposed to use HOG on depth maps. Depth maps are projected onto three orthogonal planes and the depth motion maps (DMM) are generated by accumulating global activities through entire video sequences. HOG are then computed from DMM as the representation of an action video. Another depth image-based work similar to the HOG is [74] where the histogram of oriented 4D normals (HON4D) descriptor, as a further generalization of HOG3D to four-dimensional depth videos, is proposed. HON4D descriptor calculates the histograms of oriented 4D surface normals in 4D space of time, depth, and spatial coordinates. A quantization of the 4D space is also presented. The approach in [79] is also based on the polynormal which is a cluster of neighboring hypersurface normals from a local spatiotemporal depth volume. A designed scheme aggregates the low-level polynormals in each adaptive spatiotemporal cell. The concatenation of feature vectors extracted from all spatiotemporal cells forms the final representation of depth sequences.

Jalal et al. [80] considered multifeatures from depth videos, extracting 3D human silhouettes and spatiotemporal joints values for their compact and sufficient information for HAR task.

5.2. Skeleton-Based Representations. Skeletons and joint positions are features generated from depth maps. Kinect device is popular in this representation due to its convenience of obtaining skeleton and joints. Application in Kinect v1 SDK generates 20 joints, while the later version (Kinect v2) generates 25 joints, adding 5 joints around the hands and neck (see Figure 4). We reviewed recent papers on skeleton-based representations and summarize three aspects efforts on improving the performance of skeleton-based representation.

First, skeleton model has an inherent deficiency that it always suffers the noisy skeleton problem when dealing with occlusions (see Figure 5) [76]. Features from inaccurate skeletons and joints may completely be wrong. Current approaches often solve it by combining other features that

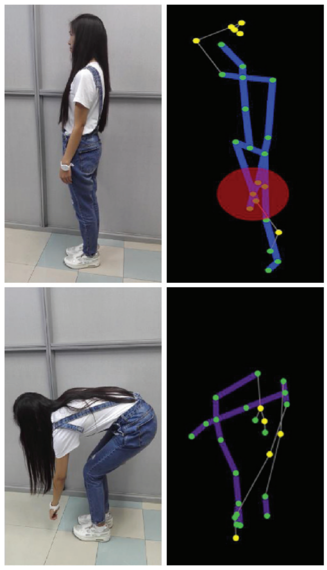


Figure 5: Noisy skeleton problem caused by self-conclusion.

robust to occlusion or alleviate occlusion problem by separating the whole skeleton into different body parts and handling them independently since not all body parts are occluded.

Second, an intuitive fact can be observed that not all skeletal joints are involved in a particular activity, and only a few active joints are meaningful and informative for a certain activity [81]. Concentrating on these active joints and abandoning the other inactive parts will generate more discriminative and robust features and are beneficial to deal with intraclass variations [82].

Finally, as an extracted feature from depth maps itself, skeleton-based representation is often combined with original depth information to form more informative and robust representation [82, 83].

Xia et al. [84] proposed a skeleton-based representation named HOJ3D, the spherical histograms of 3D locations of selected joints. After reprojected using LDA and clustered into vocabularies, the encoded features are fed to hidden Markov model (HMM) for classification. The HOJ3D is robust to view changes due to the design of the spherical coordinate system and robust skeleton estimation.

Yang and Tian [85] proposed a new type of feature named EigenJoints. 3D position differences of joints are employed to characterize three kinds of activity information including posture feature, motion feature, and offset feature. To reduce redundancy and noise, PCA is further employed and the efficient leading eigenvectors are selected. Finally, the constructed features were fed into the naïve-Bayes-nearest-neighbor (NBNN) [86] and obtained improved performance.

Wang et al. [82] indicated that using joint positions alone is insufficient to represent an action, especially for the case involving interaction with objects. Consequently, they proposed a depth-based feature called local occupancy pattern (LOP) to describe the occupancy of the neighborhood of each point, for example, the occupied space around the hand joint when lifting a cup. The local occupancy information is described by the 3D point cloud around a particular joint. Moreover, to select the active and discriminative joint feature subset (i.e., actionlet) for a particular activity, a data mining solution is leveraged and then actionlet ensemble which is linear combination of actionlets is obtained to represent each activity. Similar to actionlet, Zhu et al. [87] learned the co-occurrences of joints by designing regularization in deep LSTM (long short-term memory) RNNs (recurrent neural networks).

Shahroudy et al. [83] proposed a multimodal multipart approach for activity recognition in depth map sequences, which combines the complementary skeleton-based features LOP in [82] and depth-based features local HON4D in [74] of each part together and builds up a multimodal multipart combination. The multimodal multipart features are formulated into their framework via the proposed hierarchical mixed norm.

Chen et al. [81] proposed a skeleton-based two-level hierarchical framework. In the first layer, a part-based clustering feature vector is introduced to find out the most relevant joints and clustered them to form an initial classification. Note that the recognition task is divided into several smaller and simple tasks, which are performed within a specific cluster. It is of benefit to solving the high intraclass variance since distinct sequences of the same action are grouped into different clusters. In the second layer, only the relevant joints within specific clusters are utilized for feature extraction, which enhances the validity of the features and reduces the computational costs.

Besides depth-based features, skeleton data can be combined with other RGB features. To deal with the noisy skeleton problem, Chaaraoui et al. [88] proposed to combine skeletal and silhouette-based features using feature fusion methods. The noisy skeleton problem caused by occlusions of body part is partially elevated by the silhouette-based features. Shahroudy et al. [83] separately extracted dense trajectories features from RGB channel and 3D locations of skeleton joints from depth channel. A hierarchical feature fusion method based on structured sparsity was developed to fuse these two heterogeneous features.

6. Activity Classification Approaches

The next stage of HAR is the classification of activities that have been represented by proper feature sets extracted from images or videos. In this stage, classification algorithms give the activity label as final result. Generally speaking, most activity classification algorithms can be divided into three categories namely template-based approaches, generative models and discriminative models. Template-based approaches is a relatively simple and well accepted approach; however, it can be sometimes computationally expensive. Generative models learn a model of the joint probability $P(X,Y)$ of the inputs X and the label Y, then $P(Y|X)$ is calculated using Bayes rules and the algorithms finally picking the most likely label Y [89]. In contrast, discriminative models determine the result label directly. Typical algorithms of generative models are hidden Markov model (HMM) and dynamic Bayesian network (DBN), while support vector machine (SVM), relevance vector machine (RVM), and artificial neural network (ANN) are typical discriminative models.

6.1. Template-Based Approaches. Template-based approaches try to portray common appearance characteristics of a certain activity using various representations. These common appearance characteristics, such as 2D/3D static images/volumes or a sequence of view models, are termed as templates. Most template-based methods extract 2D/3D static templates and compare the similarity between the extracted images/volumes of test videos and the stored templates. For the classification based on a sequence of key frames, dynamic time warping (DTW) is an effective approach.

6.1.1. Template Matching. Bobick and Davis [16, 17] proposed a temporal-template-based approach. Two components, the motion-energy image (MEI) which represents the presence of motion and the motion-history image (MHI) which indicates the recency of motion, are generated for each template of an activity. In fact, the generated template images can be regarded as weighted projection of the space-time shape.

Shechtman and Irani [27, 90] constructed the 3D space–time intensity video volume template from a short training video clip. This small template is compared to every segment of same size in the test video over all three dimensions. The degree of similarity between two segments (i.e., the template and a same size video segment from the test video) is evaluated by the proposed intensity patch-based approach. It divides the segments into smaller patch units, then computes and integrates local consistency measures between those small space-time patches. This method has an impressive ability of detecting multiple different activities that occur at the same time.

Common template-based methods are unable to generate single template for each activity. They often suffer the high computational cost due to maintaining and comparing various templates. Rodriguez et al. [91] proposed to use the maximum average correlation height (MACH), which is capable of capturing intraclass variability by synthesizing a single action MACH filter for each activity class. They also generalized the MACH filter to video and vector valued data by embedding the spectral domain into a domain of Clifford algebras, building an effective approach in discriminating activities.

6.1.2. Dynamic Time Warping. Dynamic time warping (DTW) is a kind of dynamic programming algorithm for matching two sequences with variances. Rabiner and Juang [7] first developed it for speech recognition problem, representing the words as template sequence and assign matching scores for new word. DTW is also applicable to HAR problem since the human activities can be viewed as a sequence of key frames. The recognition problem is transformed to a template matching task.

Darrell and Pentland [92] proposed to build the representation of gestures using a set of learned view models. DTW algorithm is used to match the gesture template obtained from the means and variations of correlation scores between image frames and view models.

Veeraraghavan et al. [93] proposed the DTW-based nonparametric models for the gait pattern problem. They modified the DTW algorithm to include the nature of the non-Euclidean space in which the shape deformations take place. By comparing the DTW-based nonparametric and the parametric methods and applying them to the problem of gait and activity recognition, this work concluded that the DTW is more applicable than parametric modeling when there is very little domain knowledge.

Although the DTW algorithm needs a few amounts of training samples, the computational complexity increases significantly when dealing with growing activity types or those activities with high inter/intra variance, because extensive templates are needed to store those invariance.

6.2. Generative Models

6.2.1. Hidden Markov Model Approach. The recognition task is a typical evaluation problem which is one of the three hidden Markov model problems and can be solved by the forward algorithm. HMMs were initially proposed to solve the speech recognition problem [8]. Yamato et al. [94] first applied the HMM to recognize activities. Features that indicate the number of pixels in each divided mesh are obtained as observations for each frame. Then, the HMMs are trained using the observation feature vector sequences for each activity, including the initial probability of hidden states, the confusion matrix, and the transition matrix. By applying the representation mentioned above, the HAR problem (recognition of various tennis strokes) is transformed into a typical HMM evaluation problem, which can be solved using standard algorithm.

A brief summary of the deficiencies of basic HMM and several efficient extensions are presented in [95]. The basic HMM is ill-suited for modeling multiple interacting agents or body parts since it is single variable state representation, as well as those actions that have inherent hierarchical structure. Take human interaction as an example, as a kind of complex activities, it always contains more than one person in the video, to which the basic HMM is ill-suited since the standard HMM is suitable for the time structure. Another deficiency is the exponentially decayed duration model for state occupancy. This duration model has no memory of the time that has already spent on the state, which is unrealistic for activities. This is implicitly obtained from the constant state transition probability and the first-order Markov assumption, which implies that the probability of a state being observed for a certain interval of time decays exponentially with the length of the interval [96].

Previous work has proposed several variants of HMM to handle the mentioned deficiencies [95–97]. Motivated by this human interaction recognition task that have structure both in time and space (i.e., modeling activities of two or more persons), Oliver et al. [97] proposed the coupled HMM (CHMM) to model the interactions. Two HMM models are constructed for two agents and probabilities between hidden states are specified.

Flexible duration models were suggested including the hidden semi-Markov model (HSMM) and the variable

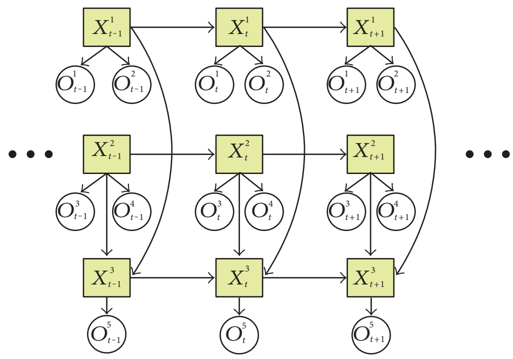


Figure 6: A typical dynamic Bayesian network [101].

transition HMMs (VT-HMM). The hidden semi-Markov model (HSMM) is a candidate approach that has explicit duration model with specific distribution. Duong et al. [98] exploited both the inherent hierarchical structure and the explicit duration model and the switching hidden semi-Markov model (S-HSMM) is introduced with two layers to represent high-level activities and atomic activities separately. Another semi-Markov model (HSMM) based work is shown in [96].

Alternatively, Ramesh and Wilpon [99] broke the implicit duration model by specifying the dependency between the transition probability and the duration. The variable transition HMMs (VT-HMMs, originally called inhomogeneous HMM in [99]) was proposed and applied in speech recognition. In VT-HMM, the transition probability of two states depends on the duration which is no longer constant. Natarajan and Nevatia [95] then presented a hierarchical variable transition HMM (HVT-HMM) based on Pamesh and Wilpon's work to recognize two-hand gestures and articulated motion of the entire body. The HVT-HMM has three layers, including a composite event layer with a single HMM representing the composite actions, a primitive event layer using a VT-HMM to represent the primitive actions, and a pose track layer with a single HMM. The pose is represented using a 23 degrees body model, including 19 degrees for joint angles, 3 degrees for direction of translation (x, y, z), and 1 degree for scale.

6.2.2. Dynamic Bayesian Networks. A dynamic Bayesian network (DBN) is a Bayesian network with the same structure unrolled in the time axis [100]. An important extension of DBN is that its state space contains more than one random variables, in contrast with the HMM that has only one single random variable. Thus, the HMM can be viewed as a simplified DBN with constrained number of random variables and fixed graph structures.

Figure 6 presents a typical DBN. Suk et al. [101] proposed this structure for two hands gesture recognition, from which we can see that there are three hidden variables. The three hidden variables represent the motion of two hands and their spatial relation, while five features including two hands' motion and the position relative to the face, as well as the spatial relation between hands are designed as observations. Then, the DBN structure is built and simplified using the first-order Markov assumptions. They proposed the DBN tailored for hands gesture recognition in contrast with the previous fixed structure of CHMM [102] which is not deemed effective for other than tight-coupled two-party interactions.

Park and Aggarwal [103] presented a hierarchical Bayesian network methodology for recognizing five two-person interactions. The proposed method first segments the body-part regions and estimates each of the body-part poses separately in the first level. Then, the individual Bayesian networks are integrated in a hierarchy to estimate the overall body poses of a person in each frame. Finally, the pose estimation results that include two-person interactions are concatenated to form a sequence with DBN algorithm.

Cherla et al. [21] indicated the contradiction for DTW between the robustness to intraclass variations and the computational complexity. Multiple templates for each activity handle the intraclass variations well but increase the computational complexity, while average templates reduce the complexity but are sensitive to intraclass variations. Cherla et al. proposed the average template with multiple feature representations to counterbalance them and achieve good performance.

6.3. Discriminative Models

6.3.1. Support Vector Machines. Support vector machines (SVMs) are typical classifiers of discriminative models and gained extensive use in HAR. Vapnik et al. [104] designed the SVM and originally used it for the problem of separating instances into two classes. It aims to find the hyperplane which maximizes the margin of two classes.

Schüldt et al. [105] combined SVM with their proposed local space-time features and applied their "local SVM approach" for HAR. A video dataset, known as the KTH dataset which had been one of the benchmarks of HAR systems, was recorded by them. The KTH dataset is introduced later in this paper (see Section 8.2.1).

Laptev et al. [50] used a nonlinear SVM with a multichannel Gaussian kernel and their SVM achieved high accuracy (91.8%) on the KTH dataset along with the HOG&HOF descriptors and local spatiotemporal bag-of-features. The well-known challenging Hollywood dataset (see Section 8.3.1) was provided and used to evaluate the proposed approach.

6.3.2. Conditional Random Fields. Conditional random fields (CRFs) are undirected graphical models that compactly represent the conditional probability of a particular label sequence Y, given a sequence of observations X. Vail et al. [106] compared the HMMs and CRFs for activity recognition. They found that the discriminatively trained CRF performed as well as or better than an HMM even when the model features are in accord with the independence assumptions of the HMM. This work pointed out a significant difference between the HMMs and CRFs: the HMMs assume that observations are independent given their labels; thus, complex features of the observation sequence will invalidate the assumption of this model and then make the HMM no longer a proper generative model. This inherent assumption of HMMs is abandoned in CRF, which conditions on the entire observation and therefore does not require any independence assumptions between the observation variables. A test was done by incorporating features which violate independence assumptions between observations (i.e., velocity thresholds in [106]) to explore the influence on both models. The result demonstrates that the CRF always outperforms the HMM, and with the increasingly severe violation of the independence assumptions, the HMM gets worse.

Natarajan and Nevatia [107] presented an approach for recognizing activities using CRF. Synthetic poses from multiple viewpoints are firstly rendered using Mocap data for known actions. Then, the poses are represented in a two-layer CRF, with observation potentials computed using shape similarity and transition potentials computed using optical flow. These basic potentials are enhanced with terms to represent spatial and temporal constraints, and the enhanced model is called the shape, flow, duration conditional random field (SFD-CRF). Single human activities as sitting down or standing up were recognized in their experiment.

Ning et al. [108] proposed a model that replaced the observation layer of a traditional random fields model with a latent pose estimator. The proposed model converted the high-dimensional observations into more compact and informative representations, and enabled transfer learning to utilize existing knowledge and data on image-to-pose relationship. This method has been shown to improve performance on the public available dataset HumanEva [109].

6.3.3. Deep Learning Architectures. Basically, the deep learning architectures can be categorized into four groups, namely deep neural networks (DNNs), convolutional neural networks (ConvNets or CNNs), recurrent neural networks (RNNs), and some emergent architectures [110].

The ConvNets is the most widely used one among the mentioned deep learning architectures. Krizhevsky et al. [9] first trained the deep ConvNets in a sufficiently large image datasets consisting of over 15 million labeled images. The impressive results lead to the extensively used of ConvNets in various pattern recognition domains [111]. Compared with traditional machine learning method and their hand-crafted features, the ConvNets can learn some representational features automatically [112]. Mo et al. [113] used ConvNets directly for feature extraction, and a multilayer perceptron is designed for the following classification.

One challenge for HAR using deep learning is how to apply it on small datasets since HAR datasets are generally smaller than what the ConvNets need. Common solutions include generating or dumpling more training instances, or converting HAR to a still image classification problem to leverage the large image dataset (e.g., ImageNet) to pretrain the ConvNets. Wang et al. [114] developed three strategies to leverage ConvNets on small training datasets. First, 3D points of depth maps are rotated to mimic different viewpoints, and WHDMMs at different temporal scales are constructed. Second, ConvNets model trained over ImageNet is adopted through transfer learning. Finally, different motion patterns are encoded into the pseudo-RGB channels with enhancement before being input to the ConvNets. On the other hand, Simonyan and Zisserm [115] leverage the large image dataset to pretrain the ConvNets. They investigated an architecture based on two separate streams (spatial and temporal), while the spatial stream contains information on appearance from still frames and is implemented using a spatial stream ConvNet. The spatial ConvNet is image classification architecture itself; thus, it is pretrained on the large image classification dataset.

The most recent research aims to further improve the performance of ConvNets by combining it with other hand-crafted features or representations. Li et al. [116] noted that the long-range dynamics information is necessary and should be modeled explicitly. Thus, they proposed a representation named $VLAD^3$, which not only captures short-term dynamics with ConvNets but also utilizes the linear dynamic systems and VLAD descriptor for medium-range and long-range dynamics. Wang et al. [117] proposed a trajectory-pooled deep-convolutional descriptor (TDD) which combined the hand-crafted local features (e.g., STIP, improved trajectories) and deep-learned features (e.g., 3D ConvNets [76, 118], two-stream ConvNets [115]). The proposed TDD integrates the advantages of these two features and adopts the state-of-the-art improved trajectories and two-stream ConvNets.

Unlike ConvNets, DNNs still use hand-crafted features instead of automatically learning features by deep networks from raw data. Berlin and John [119] used Harris corner-based interest points and histogram-based features as input. The proposed deep neural network with stacked auto encoders are used to recognize human-human interactions. Huang et al. [120] learned Lie group features (i.e., one of the skeletal data representations that are learned by manifold-based approaches) by incorporating a Lie group structure into a deep network architecture.

RNNs are designed for sequential information and have been explored successfully in speech recognition and natural language processing [121, 122]. Activity itself is a kind of time-series data and it is a natural thought to use RNNs for activity recognition.

Among various RNNs architectures, the long short-term memory (LSTM) is the most popular one as it is able to maintain observations in memory for extended periods of time [123]. As an initial study for activity recognition, a LSTM network was utilized to classify activities in soccer videos [124]. Then, further research [123] explicitly demonstrated the robustness of LSTM even as experimental conditions deteriorate and indicated its potential for robust real-world recognition. Veeriah et al. [125] extended the LSTM to differential recurrent neural networks (RNNs). By computing the different orders of derivative of state which is sensitive to the spatiotemporal structure, the salient spatiotemporal representations of actions are learned, while in contrast, the conventional LSTM does not capture salient dynamic patterns of activity.

In addition to videos, RNNs can also be applied to skeleton data for activity recognition. Du et al. [126] proposed a hierarchical RNNs structure for skeleton-based recognition.

The human skeleton from Kinect are divided into five parts and are fed into subnets separately. Representations from subnets are hierarchically fused into a higher layer and finally fed into a single-layer perceptron, whose temporally accumulated output is the final decision.

A detailed taxonomy about the representation, classification methods, and the used datasets of the introduced works in this review are presented in Table 1.

7. Human Tracking Approaches

Besides the activity classification approaches, another critical research area is the human tracking approach, which is widely concerned in video surveillance systems for suspicious behavior detection. Human tracking is performed to locate a person along the video sequence over a time period, and then the resultant trajectories of people are further processed by expert surveillance systems for analyzing human behaviors and identifying potential unsafe or abnormal situations [127]. In this section, we briefly review recent literatures of two dominant approaches, namely kernel-based tracking and filtering-based tracking.

7.1. Filter-Based Tracking. Filtering is one of the widely used approaches for tracking, and the representative Kalman filter (KF) [128] and particle filter (PF) [129] are two commonly used classic filtering techniques.

KF is a state estimate method based on linear dynamical systems that are perturbed by Gaussian noise [130]. Patel and Thakore utilized traditional KF to track moving objects, in both the indoor and outdoor places. Vijay and Johnson [131] also utilized traditional KF for tracking moving objects such as car or human. However, the tested scenarios of these cases are relatively spacious and thus seldom occlusion occur. Despite the good results that are achieved by the KF-based method, it is strictly constrained with effective foreground segmentation, and its ability is limited when handling the occlusion cases. Arroyo et al. [127] combined Kalman filtering with a linear sum assignment problem (LSAP). To deal with the occlusion problem, visual appearance information is used with image descriptors of GCH (global color histogram), LBP (local binary pattern), and HOG (histogram of oriented gradients) representing the color, texture, and gradient information, respectively.

Particle filter, or sequential Monte Carlo method [132], is another typical filtering method for tracking. PF is a conditional density propagation method that is utilized to deal with non-Gaussian distributions and multimodality cases [130]. Ali et al. [133] combined a head detector and particle filter for tracking multiple people in high-density crowds. Zhou et al. [130] presented a spatiotemporal motion energy particle filter for human tracking, which fuses the local features of colour histograms as well as the spatiotemporal motion energy. The proposed particle filter-based tracker achieved robustness to illumination changes and temporal occlusions through using these features, as the motion energy contains the dynamic characteristics of the targeted human. As a specific branch of particle filter research, the sequential Monte Carlo implementation of the probability hypothesis density (PHD) filter, known as the particle PHD filter, is well developed for solving multiple human tracking problems. A series of research have been conducted by Feng et al. in [134–138].

7.2. Kernel-Based Tracking. Kernel-based tracking [139] or mean shift tracking [140] tracks the object (human) by computing the motion of one or more spatially weighted color histograms (i.e., single kernel/multiple kernels) from the current frame to next frame based on an iteratively mean-shift procedure. The kernel-based approach has fast convergence speed and low computation requirement inherited from the efficient mean shift procedure [141].

Traditional kernel-based tracking used symmetric constant kernel, and it tends to encounter problems of object scale and object orientation variation, as well as the object shape deformation. Research was conducted concerning these problems. Liu et al. [142] presented a kernel-based tracking algorithm based on eigenshape kernel. Yilmaz [143] introduced a kernel-based tracking algorithm based on asymmetric kernel for the first time. This kernel uses the initial region inside the outline of the target as kernel template and generates a precise tracking contour of the object. Yuan-ming et al. [144] noticed the shortage of the fixed asymmetric kernel. They combined the contour evolution technology with the mean shift and proposed an enhanced mean shift tracking algorithm based on evolutive asymmetric kernel. Liu et al. [145] presented an adaptive shape kernel-based mean shift tracker. Shape of the adaptive kernel is reconstructed from the low-dimensional shape space obtained by nonlinear manifold learning technique to the high-dimensional shape space, aiming to be adaptive to the object shape.

Early literatures reported tracking methods using single kernel scheme. However, the single kernel-based tracking could fail when the human is concluded, that is, the object could be lost or mismatch due to the partial observation. Thus, multiple-kernel tracking is adopted in most cases of recent researches. Lee et al. [146] evaluated two kernel and four kernel schemes [147] and presented a similar two and four kernal evaluation. Chu et al. [148] proposed to utilize projected gradient to facilitate multiple-kernel tracking in finding the best match under predefined constraints. The occlusion is managed by employing adaptive weights, that is, decreasing the importance of the kernel being occluded whilst enhancing the ones which are well-observed. Hou et al. [149] integrated the deformable part model (DPM) and designed multiple kernels, each of which corresponds to a part model of a DPM-detected human.

8. Representative Datasets in HAR

Public datasets could be used to compare different approaches in the same standards therefore accelerate the development of HAR methods. In this section, several representative datasets are reviewed, organized as a three-level category mentioned in the beginning of this review (i.e., action primitive level, action/activity level, and interaction level). There have been a published good survey [4] which presents

Table 3: Overview of representative datasets.

Dataset	Modality	Level	Year	References	Web pages	Activity category
RGBD-HuDaAct	RGB-D	Interaction level	2013	[163]	http://adsc.illinois.edu/sites/default/files/files/ADSC-RGBD-dataset-download-instructions.pdf	12 classes: eat meal, drink water, mop floor, and so forth
Hollywood	RGB	Interaction level	2008	[50]	http://www.di.ens.fr/~laptev/download.html#actionclassification	8 classes: answer phone, hug person, kiss, and so forth
Hollywood-2	RGB	Interaction level	2009	[158]	http://www.di.ens.fr/~laptev/download.html#actionclassification	12 classes: answer phone, driving a car, fight, and so forth
UCF sports	RGB	Interaction level	2008	[91]	http://crcv.ucf.edu/data/UCF_Sports_Action.php	10 classes: golf swing, diving, lifting, and so forth
KTH	RGB	Activity/action level	2004	[105]	http://www.nada.kth.se/cvap/actions/	6 classes: walking, jogging, running, and so forth
Weizmann	RGB	Activity/action level	2005	[5]	http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html	10 classes: run, walk, bend, jumping-jack, and so forth
NTU-MSR	RGB-D	Action primitive level	2013	[152]	http://web.cs.ucla.edu/~zhou.ren/	10 classes: it contains 10 different gestures.
MSRC-Gesture	RGB-D	Action primitive level	2012	[153]	http://research.microsoft.com/en-us/um/cambridge/projects/msrc12/	12 classes: it contains 12 different gestures.
MSR DailyAction3D	RGB-D	Interaction level	2012	[160]	http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm	16 classes: call cellphone, use laptop, walk, and so forth
MSR Action3D	Depth	Activity/action level	2010	[76]	http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm	20 classes: high arm wave, hand clap, jogging, and so forth

the available important public datasets; however, it mainly focused on the conventional RGB-based datasets and missed current depth-based datasets. Thus, several important benchmark depth or RGB-D datasets are also reviewed in this section, with an overview of them (Table 3).

8.1. Action Primitive Level Datasets. While action primitives often act as components of high level human activities (e.g., the action primitives are served as a layer in hierarchical HMM to recognize activities [95] or interactions [97]), some typical and meaningful action primitives, such as poses and gestures [150], gait pattern [151], are studied as separate topics. These topics aroused wide research interest due to their importance in applications such as human-computer interaction and health care. Here, we present two recent gesture dataset based on RGB-D as the representative dataset in this level.

8.1.1. NTU-MSR Kinect Hand Gesture Dataset (2013). The NTU-MSR Kinect hand gesture dataset [152] is considered as an action primitive level since it is developed for gesture recognition. Gestures in it were collected by Kinect, and each of them consists of a color image and the corresponding depth map. Totally, 1000 cases of 10 gestures were collected by 10 subjects, and each gesture was performed 10 times by a single subject in different poses. The dataset is claimed as a challenging real-life dataset due to their cluttered backgrounds. Besides, for each gesture, the subject poses with variations in hand orientation, scale, articulation, and so forth.

8.1.2. MSRC-Kinect Gesture Dataset (2012). The MSRC-Kinect gesture dataset [153] is another typical action primitive level dataset, in which large amounts of limb level movements (e.g., karate kicking forwards with right leg) were recorded. There are totally 6244 instances of 12 gestures performed by 30 people, collected by Kinect. Positions of 20 tracked joints are provided as well.

8.2. Action/Activity Level Datasets. According to our definition, action/activity is middle level human activity without any human-human or human-object interactions. We first review two classic datasets, namely KTH human activity dataset and Weizmann human activity dataset. Though these two datasets have gradually faded out of state-of-the-art and are considered as easy tasks (e.g., 100% accuracy for Weizmann in [18, 25, 95]), they did play important roles in the history and act as benchmarks in earlier HAR works. Then, the well-known benchmark dataset for depth-based approaches, MSR Action3D dataset, is introduced next.

8.2.1. KTH Activity Dataset (2004). The KTH dataset [105] is one of the most frequently cited datasets. It contains 6 activities (walking, jogging, running, boxing, hand waving, and hand clapping) performed by 25 subjects in controlled sceneries including outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. One important factor in their success is the high intraclass variation in it which is one of the criteria for evaluation algorithms. Although the videos were still taken using static cameras, the high variation details, such as various scenarios and actors' clothes, as well as the different viewpoints, make itself a fair and convincing datasets for comparison. Most of the collected human activities in it were performed by a single person without any human-object interaction; thus, it is categorized in the activity/action level.

8.2.2. Weizmann Activity Dataset (2005). The Weizmann activity dataset [5] was created by the Weizmann Institute of Science (Israel) in 2005. The Weizmann dataset consists of 10 natural actions (running, walking, skipping, bending, jumping-jack, galloping-sideways, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, waving-two-hands, and waving-one-hand) with 10 subjects. Totally, 90 video sequences in a low resolution of $180^{*}144$, 50 fps were recorded using a fixed camera and a simple background. To address the robustness of the proposed algorithm in [5], ten additional video sequences of people walking in various complicated scenarios in front of different nonuniform backgrounds were collected. Similar to the KTH dataset, most human activities in Weizmann were performed by a single person without any human-object interaction; thus, it is categorized in the activity/action level.

8.2.3. MSR Action3D Dataset (2010). The MSR Action3D dataset [76] is widely used as the benchmark for depth-based HAR approaches. Depth maps of 20 activity classes performed by 10 subjects are provided in it (high arm waving, horizontal arm waving, hammering, hand catching, forward punching, high throwing, drawing cross, drawing tick, drawing circle, clapping hand, waving two hand, side-boxing, bending, forward kicking, side kicking, jogging, tennis swing, tennis serve, golf swing, pickup, and throw). MSR Action3D is a pure depth datasets without any color images in it.

8.3. Interaction Level Datasets. Interaction level datasets are relatively difficult tasks. Due to the human or human-object interactions, interaction level human activities are more realistic and abound in various scenarios such as sport events [91], video surveillance, and different movie scenes [50]. In this section, we review two conventional RGB datasets (i.e., Hollywood human activity dataset and UCF sports human activity dataset) and a RGB-D dataset (i.e., MSR DailyActivity3D dataset). Designed to cover indoor daily activities, MSR DailyActivity3D dataset [160] is more challenging and involves more human-object interactions compared to MSR Action3D [82].

8.3.1. Hollywood Human Activity Dataset (2008 and 2009). Another well-known interaction level dataset is the Hollywood human activity dataset [50, 158]. As a representative of realistic activity dataset, the Hollywood dataset is introduced here as a challenging task compared to previous datasets due to its frequently moved camera viewpoints, occlusions, and dynamic backgrounds with seldom provided information [1]. The initial version published in 2008 [50] contains approximately 663 video samples (233 samples in automatic training set, 219 samples in clean training set, and 211 samples in test set) of eight actions (answering phone, getting out of car, hugging, handshaking, kissing, sitting down, sitting up, and standing up) from 32 movies. Recognition of natural human activities in diverse and realistic video settings, which can be tested on this dataset, was discussed in [50]. Then, the extended Hollywood dataset was created in 2009 [158], involving four additional activities (driving a car, eating, fighting, and running) and more samples for each class, totally, 3669 video clips from 69 movies. Both human interaction (e.g., kissing, fighting) and human-object interactions (e.g., answering phone, driving a car) are included. Marszalek et al. [158] exploited the relationship between context of natural dynamic scenes and human activities in video based on this extended Hollywood dataset.

8.3.2. UCF Sports Dataset (2007). The UCF sports dataset [91] is a specific interaction level dataset focused on various sports activities from television broadcasts. It is one of the datasets collected by Computer Vision Lab, University of Central Florida. There are over 200 video sequences in this dataset, covering 9 sport activities including diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging a basketball bat, and pole vaulting. While it covers only 9 human activities in sports scenes, it is still a challenging task for recognition due to its unconstrained environment and abound intraclass variability.

8.3.3. MSR DailyAction3D Dataset (2012). The MSR DailyActivity3D dataset [160] is an interactive level dataset captured by Kinect device. In contrast with the previous MSR Action3D, this dataset provides three types of data including depth maps, skeleton joint positions, and RGB video. 16 activity classes performed by 10 subjects (drinking, eating, reading book, calling cellphone, writing on a paper, using laptop, using vacuum cleaner, cheering up, sitting still, tossing paper, playing game, lying down on sofa, walking, playing guitar, standing up, and sitting down) are recorded in it.

9. Conclusions and Future Direction

Human activity recognition remains to be an important problem in computer vision. HAR is the basis for many applications such as video surveillance, health care, and human-computer interaction. Methodologies and technologies have made tremendous development in the past decades and have kept developing up to date. However, challenges still exist when facing realistic sceneries, in addition to the inherent intraclass variation and interclass similarity problem.

In this review, we divided human activities into three levels including action primitives, actions/activities, and interactions. We have summarized the classic and representative approaches to activity representation and classification, as well as some benchmark datasets in different levels. For representation approaches, we roughly sorted out the research trajectory from global representations to local representations and recent depth-based representations. The literatures were reviewed in this order. State-of-the-art approaches, especially those depth-based representations, were discussed, aiming to cover the recent development in HAR domain. As the next step, classification methods play important roles and prompt the advance of HAR. We categorized classification approaches into template-matching methods, discriminative models, and generative models. Totally, 7 types of method from the classic DTW to the newest deep learning were summarized. For human tracking approaches, two categories are considered namely filter-based and kernel-based human tracking. Finally, 7 datasets were introduced, covering different levels from primitive level to interaction level, ranging from classic datasets to recent benchmark for depth-based methods.

Though recent HAR approaches have achieved great success up to now, applying current HAR approaches in real-world systems or applications is still nontrivial. Three future directions are recommended to be considered and further explored.

First, current well-performed approaches are mostly hard to be implemented in real time or applied to wearable devices, as they are subject to constrained computing power. It is difficult for computational constrained systems to achieve comparable performances of those offline approaches. Existing work utilized additional inertial sensors to assist in recognizing, or developed microchips, for embedded devices. Besides these hardware-oriented solutions, from a computer vision perspective, more efficient descriptor extracting methods and classification approaches are expected to train recognition models fast, even in real time. Another possible way is to degrade quality of input image and strike a balance among input information, algorithm efficiency, and recognizing rate. For example, utilizing depth maps as inputs and abandoning color information are ways of degrading quality.

Second, many of the recognition tasks are solved case by case, for both the benchmark datasets and the recognition methods. The future direction of research is obviously encouraged to unite various datasets as a large, complex, and complete one. Though every dataset may act as benchmark in its specific domain, uniting all of them triggers more effective and general algorithms which are more close to real-world occasions. For example, recent deep learning is reported to perform better in a four-dataset-combined larger datasets [114]. Another promising direction is to explore an evaluation criterion which enables comparisons among wide variety of recognition methods. Specifically, several vital measuring indexes are defined and weighted according to specific task, evaluating methods by measuring indexes such as recognition rate, efficiency, robustness, number, and level of recognizable activities.

Third, mainstream recognition system remains in a relatively low level comparing with those higher level behaviors. Ideally, the system should be able to tell the behavior "having a meeting" rather than lots of people sitting and talking, or even more difficult, concluding that a person hurried to catch a bus rather than just recognizing "running." Activities are analogous to the words consisting behavior languages. Analyzing logical and semantic relations between behaviors and activities is an important aspect, which can be learned by transferring from Natural language processing (NLP) techniques. Another conceivable direction is to derive additional features from contextual information. Though this direction has been largely exploited, current approaches usually introduce all the possible contextual variables without screening. This practice not only reduces the efficiency but also affects the accuracy. Thus, dynamically and reasonably choosing contextual information is a future good topic to be discussed.

Finally, though recent deep learning approaches achieve remarkable performance, a conjoint ConvNets + LSTM architecture is expected for activity video analysis in the future. On the one hand, ConvNets are spatial extension of conventional neural networks and exhibit its advantage in the image classification tasks. This structure captures the spatial correlation characteristics, however, ignores the temporal dependencies of the interframe content for activity dynamics modeling. On the other hand, LSTM as a representative kind of RNN, is able to model the temporal or sequence information, which makes up the temporal shortage of ConvNets. LSTM is currently used in accelerometer-based recognition, skeleton-based activity recognition, or one-dimensional signal processing, but has not been widely concerned in combination with ConvNets for two-dimensional video activity recognition, which we believe is a promising direction in the future.

Conflicts of Interest

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

Acknowledgments

This research is supported by the National Natural Science Foundation of China (no. 61602430, no. 61672475, and no. 61402428); major projects of Shandong Province (no. 2015ZDZX05002); Qingdao Science and Technology Development Plan (no. 16-5-1-13-jch); and The Aoshan Innovation Project in Science and Technology of Qingdao National Laboratory for Marine Science and Technology (no. 2016ASKJ07).

References

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, pp. 976–990, 2010.

[2] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," *ACM Computing Surveys*, vol. 43, p. 16, 2011.

[3] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, pp. 90–126, 2006.

[4] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, pp. 633–659, 2013.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pp. 1395–1402, Beijing, China, 2005.

[6] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proceedings Ninth IEEE International Conference on Computer Vision*, vol. 1, pp. 432–439, Nice, France, 2003.

[7] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, 1993.

[8] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, pp. 4–16, Prentice Hall, Upper Saddle River, New Jersey, 1986.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, Lake Tahoe, Nevada, 2012.

[10] A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 726–733, Nice, France, 2003.

[11] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780–785, 1997.

[12] D. Koller, J. Weber, T. Huang et al., "Towards robust automatic traffic scene analysis in real-time, in: pattern recognition," in *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, pp. 126–131, Jerusalem, Israel, 1994.

[13] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, pp. 246–252, Fort Collins, CO, USA, 1999.

[14] A. Veeraraghavan, A. R. Chowdhury, and R. Chellappa, "Role of shape and kinematics in human movement analysis," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, pp. I–730, Washington, DC, USA, 2004.

[15] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury, "The function space of an activity," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 959–968, New York, NY, USA, 2006.

[16] A. Bobick and J. Davis, "An appearance-based representation of action," in *Proceedings of 13th International Conference on Pattern Recognition*, pp. 307–312, Vienna, Austria, 1996.

[17] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 257–267, 2001.

[18] N. Ikizler and P. Duygulu, "Human action recognition using distribution of oriented rectangular patches," in *Human Motion–Understanding Modelling Capture and Animation*, pp. 271–284, Springer, Rio de Janeiro, Brazil, 2007.

[19] G. Xu and F. Huang, "Viewpoint insensitive action recognition using envelop shape," in *Computer Vision–Asian Conference on Computer Vision 2007*, pp. 477–486, Springer, Tokyo, Japan, 2007.

[20] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, pp. 249–257, 2006.

[21] S. Cherla, K. Kulkarni, A. Kale, and V. Ramasubramanian, "Towards fast, view-invariant human action recognition," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8, Anchorage, AK, USA, 2008.

[22] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81 Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, pp. 674–679, Vancouver, British Columbia, Canada, 1981.

[23] J. Shi and C. Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600, Seattle, WA, USA, 1994.

[24] X. Lu, Q. Liu, and S. Oe, "Recognizing non-rigid human actions using joints tracking in space-time," in *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004*, pp. 620–624, Las Vegas, NV, USA, 2004.

[25] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *ECCV '08 Proceedings of the 10th European Conference on Computer Vision: Part I*, pp. 548–561, Springer, Amsterdam, The Netherlands, 2008.

[26] C. Achard, X. Qu, A. Mokhber, and M. Milgram, "A novel approach for recognition of human actions with semi-global features," *Machine Vision and Applications*, vol. 19, pp. 27–34, 2008.

[27] E. Shechtman and M. Irani, "Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2045–2056, 2007.

[28] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, 2007.

[29] S. Kumari and S. K. Mitra, "Human action recognition using DFT," in *2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pp. 239–242, Hubli, Karnataka, India, 2011.

[30] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Conference: Proceedings of the British Machine Vision Conference 2008*, pp. 271–275, Leeds, United Kingdom, 2008.

[31] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice, arXiv Prepr," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, Elsevier, Amsterdam, The Netherlands, 2016, http://arxiv.org/abs/1405.4506.

[32] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," *Computer Vision–Asian Conference on Computer Vision–ECCV 2014*, pp. 581–595, Springer, Zurich, 2014.

[33] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj, "Beyond gaussian pyramid: multi-skip feature stacking for action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 204–212, Boston, MA, USA, 2015.

[34] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, p. 50, Manchester, UK, 1988.

[35] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, pp. 83–105, 2001.

[36] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 3, pp. 710–719, 2005.

[37] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and*

Performance Evaluation of Tracking and Surveillance, pp. 65–72, Beijing, China, 2005.

[38] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pp. 166–173, Beijing, China, 2005.

[39] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Computer Vision—European Conference on Computer Vision 2008*, pp. 650–663, Springer, Marseille, France, 2008.

[40] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pp. 1–511, Kauai, HI, USA, 2001.

[41] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, pp. 107–123, 2005.

[42] O. Oshin, A. Gilbert, J. Illingworth, and R. Bowden, "Spatio-temporal feature recogntion using randomised ferns," in *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis-MLVMA'08*, pp. 1–12, Marseille, France, 2008.

[43] M. Ozuysal, P. Fua, and V. Lepetit, "Fast keypoint recognition in ten lines of code," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, MN, USA, 2007.

[44] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pp. 1150–1157, Kerkyra, Greece, 1999.

[45] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[46] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 357–360, Augsburg, Germany, 2007.

[47] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, 2008.

[48] 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, 2005.

[49] W. L. Lu and J. J. Little, "Simultaneous tracking and action recognition using the pca-hog descriptor," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, Quebec, Canada, p. 6, 2006.

[50] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, AK, USA, 2008.

[51] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3169–3176, Colorado Springs, CO, USA, 2011.

[52] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, pp. 363–370, Springer, Halmstad, Sweden, 2003.

[53] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *2013 IEEE International Conference on Computer Vision*, pp. 3551–3558, Sydney, Australia, 2013.

[54] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[55] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[56] X. Wang, L. Wang, and Y. Qiao, "A comparative study of encoding, pooling and normalization methods for action recognition," in *Computer Vision–Asian Conference on Computer Vision 2012*, pp. 572–585, Springer, Sydney, Australia, 2012.

[57] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference*, p. 8, University of Dundee, 2011.

[58] X. Zhen and L. Shao, "Action recognition via spatio-temporal local features: a comprehensive study," *Image and Vision Computing*, vol. 50, pp. 1–13, 2016.

[59] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 1470–1477, Nice, France, 2003.

[60] J. C. V. Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Computer Vision–European Conference on Computer 2008*, pp. 696–709, Springer, Marseille, France, 2008.

[61] J. C. V. Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1271–1283, 2010.

[62] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1794–1801, Miami, FL, USA, 2009.

[63] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advance Neural Information Processing Systems*, pp. 2223–2231, Vancouver, British Columbia, Canada, 2009.

[64] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367, San Francisco, CA, USA, 2010.

[65] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Computer Vision–European Conference on Computer Vision 2010*, pp. 143–156, Springer, Heraklion, Crete, Greece, 2010.

[66] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 215–223, Fort Lauderdale, USA, 2011.

[67] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311, San Francisco, CA, USA, 2010.

[68] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into

compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1704–1716, 2012.

[69] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Computer Vision–European Conference on Computer Vision 2010*, pp. 141–154, Springer, Heraklion, Crete, Greece, 2010.

[70] K. Yu and T. Zhang, "Improved local coordinate coding using local tangents," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1215–1222, Haifa, Israel, 2010.

[71] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *2011 International Conference on Computer Vision*, pp. 2486–2493, Barcelona, Spain, 2011.

[72] Y. Huang, K. Huang, Y. Yu, and T. Tan, "Salient coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference*, pp. 1753–1760, Colorado Springs, CO, USA, 2011.

[73] Z. Wu, Y. Huang, L. Wang, and T. Tan, "Group encoding of local features in image classification," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 1505–1508, Tsukuba, Japan, 2012.

[74] O. Oreifej and Z. Liu, "Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, Portland, OR, USA, 2013.

[75] J. Shotton, T. Sharp, A. Kipman et al., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, pp. 116–124, 2013.

[76] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern*, pp. 9–14, San Francisco, CA, USA, 2010.

[77] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, "Combing rgb and depth map features for human activity recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–4, Hollywood, CA, USA, 2012.

[78] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 1057–1060, Nara, Japan, 2012.

[79] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1028–1039, 2017.

[80] A. Jalal, S. Kamal, and D. Kim, "A depth video-based human detection and activity recognition using multi-features and embedded hidden Markov models for health care monitoring systems," *International Journal of Interactive Multimedia and Artificial Inteligence*, vol. 4, no. 4, p. 54, 2017.

[81] H. Chen, G. Wang, J. H. Xue, and L. He, "A novel hierarchical framework for human action recognition," *Pattern Recognition*, vol. 55, pp. 148–159, 2016.

[82] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 914–927, 2014.

[83] A. Shahroudy, T. T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2123–2129, 2016.

[84] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, Providence, RI, USA, 2012.

[85] X. Yang and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 14–19, Providence, RI, USA, 2012.

[86] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, AK, USA, 2008.

[87] W. Zhu, C. Lan, J. Xing et al., *Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep Lstm Networks, arXiv Prepr*, vol. 2, AAAI, p. 8, 2016, http://arxiv.org/abs/1603.07772.

[88] A. Chaaraoui, J. Padilla-Lopez, and F. Flórez-Revuelta, "Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices," in *2013 IEEE International Conference on Computer Vision Workshops*, pp. 91–97, Sydney, Australia, 2013.

[89] A. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, p. 841, 2002.

[90] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 405–412, San Diego, CA, USA, 2005.

[91] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, AK, USA, 2008.

[92] T. Darrell and A. Pentland, "Space-time gestures," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 335–340, New York, NY, USA, 1993.

[93] A. Veeraraghavan, A. K. Roy-Chowdhury, and R. Chellappa, "Matching shape sequences in video with applications in human movement analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1896–1909, 2005.

[94] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 379–385, Champaign, IL, USA, 1992.

[95] P. Natarajan and R. Nevatia, "Online, real-time tracking and recognition of human actions," in *2008 IEEE Workshop on Motion and Video Computing*, pp. 1–8, Copper Mountain, CO, USA, 2008.

[96] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden markov models," in *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 1455–1462, Nice, France, 2003.

[97] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 831–843, 2000.

[98] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the

switching hidden semi-markov model," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern*, pp. 838–845, San Diego, CA, USA, 2005.

[99] P. Ramesh and J. G. Wilpon, "Modeling state durations in hidden Markov models for automatic speech recognition," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 381–384, San Francisco, CA, USA, 1992.

[100] Y. Luo, T. D. Wu, and J. N. Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks," *Computer Vision and Image Understanding*, vol. 92, pp. 196–216, 2003.

[101] H. I. Suk, B. K. Sin, and S. W. Lee, "Hand gesture recognition based on dynamic Bayesian network framework," *Pattern Recognition*, vol. 43, pp. 3059–3072, 2010.

[102] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 994–999, San Juan, Puerto Rico, USA, 1997.

[103] S. Park and J. K. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Multimedia System*, vol. 10, pp. 164–179, 2004.

[104] V. Vapnik, S. E. Golowich, and A. Smola, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems*, vol. 9, pp. 841–848, Vancouver, British Columbia, Canada, 1996.

[105] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Pattern Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, pp. 32–36, Cambridge, UK, 2004.

[106] D. L. Vail, M. M. Veloso, and J. D. Lafferty, "Conditional random fields for activity recognition," in *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, p. 235, Honolulu, Hawaii, 2007.

[107] P. Natarajan and R. Nevatia, "View and scale invariant action recognition using multiview shape-flow models," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Anchorage, AK, USA, 2008.

[108] H. Ning, W. Xu, Y. Gong, and T. Huang, "Latent pose estimator for continuous action recognition," in *Computer Vision–European Conference on Computer Vision 2008*, pp. 419–433, Springer, Marseille, France, 2008.

[109] L. Sigal and M. J. Black, *Humaneva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion, Brown Univertsity TR*, p. 120, 2006.

[110] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 17, 2016.

[111] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[112] G. Luo, S. Dong, K. Wang, and H. Zhang, "Cardiac left ventricular volumes prediction method based on atlas location and deep learning," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1604–1610, Shenzhen, China, 2016.

[113] L. Mo, F. Li, Y. Zhu, and A. Huang, "Human physical activity recognition based on computer vision with deep learning model," in *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pp. 1–6, Taipei, Taiwan, 2016.

[114] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 498–509, 2016.

[115] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, pp. 568–576, Montreal, Quebec, Canada, 2014.

[116] Y. Li, W. Li, V. Mahadevan, and N. Vasconcelos, "Vlad3: encoding dynamics of deep features for action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1951–1960, Las Vegas, NV, USA, 2016.

[117] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4305–4314, Boston, MA, USA, 2015.

[118] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 221–231, 2013.

[119] S. J. Berlin and M. John, "Human interaction recognition through deep learning network," in *2016 IEEE International Carnahan Conference on Security Technology (ICCST)*, pp. 1–4, Orlando, FL, USA, 2016.

[120] Z. Huang, C. Wan, T. Probst, and L. V. Gool, *Deep Learning on lie Groups for Skeleton-Based Action Recognition, arXiv Prepr*, Cornell University Library, Ithaca, NY, USA, 2016, http://arxiv.org/abs/1612.05877.

[121] R. Kiros, Y. Zhu, R. R. Salakhutdinov et al., "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, pp. 3294–3302, Montreal, Quebec, Canada, 2015.

[122] J. Li, M.-T. Luong, and D. Jurafsky, *A Hierarchical Neural Autoencoder for Paragraphs and Documents, arXiv Prepr*, Cornell University Library, Ithaca, NY, USA, 2015, http://arxiv.org/abs/1506.01057.

[123] A. Grushin, D. D. Monner, J. A. Reggia, and A. Mishra, "Robust human action recognition via long short-term memory," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Dallas, TX, USA, 2013.

[124] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Action classification in soccer videos with long short-term memory recurrent neural networks," in *International Conference Artificial Neural Networks*, pp. 154–159, Thessaloniki, Greece, 2010.

[125] V. Veeriah, N. Zhuang, and G. J. Qi, "Differential recurrent neural networks for action recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4041–4049, Santiago, Chile, 2015.

[126] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, Boston, MA, USA, 2015.

[127] R. Arroyo, J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazán, "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls," *Expert Systems with Applications*, vol. 42, pp. 7991–8005, 2015.

[128] G. Welch and G. Bishop, *An Introduction to the Kalman Filter*, University of North Carolina at Chapel Hill Chapel Hill NC, Chapel Hill, North Carolina, USA, 1995.

[129] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *IEE Proceedings F - Radar and Signal Processing*, pp. 107–113, London, UK, 1993.

[130] H. Zhou, M. Fei, A. Sadka, Y. Zhang, and X. Li, "Adaptive fusion of particle filtering and spatio-temporal motion energy for human tracking," *Pattern Recognition*, vol. 47, pp. 3552–3567, 2014.

[131] A. A. Vijay and A. K. Johnson, "An integrated system for tracking and recognition using Kalman filter," in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 1065–1069, Kanyakumari, India, 2014.

[132] P. Sarkar, *Sequential Monte Carlo Methods in Practice*, Taylor & Francis, Oxfordshire, UK, 2003.

[133] I. Ali and M. N. Dailey, "Multiple human tracking in high-density crowds," *Image and Vision Computing*, vol. 30, pp. 966–977, 2012.

[134] P. Feng, W. Wang, S. Dlay, S. M. Naqvi, and J. Chambers, "Social force model-based MCMC-OCSVM particle PHD filter for multiple human tracking," *IEEE Transactions on Multimedia*, vol. 19, pp. 725–739, 2017.

[135] P. Feng, W. Wang, S. M. Naqvi, and J. Chambers, "Adaptive retrodiction particle PHD filter for multiple human tracking," *IEEE Signal Processing Letters*, vol. 23, pp. 1592–1596, 2016.

[136] P. Feng, W. Wang, S. M. Naqvi, S. Dlay, and J. A. Chambers, "Social force model aided robust particle PHD filter for multiple human tracking," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4398–4402, Shanghai, China, 2016.

[137] P. Feng, M. Yu, S. M. Naqvi, W. Wang, and J. A. Chambers, "A robust student's-t distribution PHD filter with OCSVM updating for multiple human tracking," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 2396–2400, Nice, France, 2015.

[138] P. Feng, *Enhanced Particle PHD Filtering for Multiple Human Tracking, School of Electrical and Electronic Engineering, Newcastle University*, Newcastle University, Newcastle upon Tyne, UK, 2016.

[139] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 564–577, 2003.

[140] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[141] L. Hou, W. Wan, K. Han, R. Muhammad, and M. Yang, "Human detection and tracking over camera networks: a review," in *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 574–580, Shanghai, China, 2016.

[142] C. Liu, C. Hu, and J. K. Aggarwal, "Eigenshape kernel based mean shift for human tracking," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1809–1816, Barcelona, Spain, 2011.

[143] A. Yilmaz, "Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, Minneapolis, MN, USA, 2007.

[144] D. Yuan-ming, W. Wei, L. Yi-ning, and Z. Guo-xuan, "Enhanced mean shift tracking algorithm based on evolutive asymmetric kernel," in *2011 International Conference on Multimedia Technology*, pp. 5394–5398, Hangzhou, China, 2011.

[145] C. Liu, Y. Wang, and S. Gao, "Adaptive shape kernel-based mean shift tracker in robot vision system," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 6040232, 8 pages, 2016.

[146] K. H. Lee, J. N. Hwang, G. Okopal, and J. Pitton, "Ground-moving-platform-based human tracking using visual SLAM and constrained multiple kernels," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, pp. 3602–3612, 2016.

[147] Z. Tang, J. N. Hwang, Y. S. Lin, and J. H. Chuang, "Multiple-kernel adaptive segmentation and tracking (MAST) for robust object tracking," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1115–1119, Shanghai, China, 2016.

[148] C. T. Chu, J. N. Hwang, H. I. Pai, and K. M. Lan, "Tracking human under occlusion based on adaptive multiple kernels with projected gradients," *IEEE Transactions on Multimedia*, vol. 15, pp. 1602–1615, 2013.

[149] L. Hou, W. Wan, K. H. Lee, J. N. Hwang, G. Okopal, and J. Pitton, "Robust human tracking based on DPM constrained multiple-kernel from a moving camera," *Journal of Signal Processing Systems*, vol. 86, pp. 27–39, 2017.

[150] S. Mitra and T. Acharya, "Gesture recognition: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, pp. 311–324, 2007.

[151] S. Mulroy, J. Gronley, W. Weiss, C. Newsam, and J. Perry, "Use of cluster analysis for gait pattern classification of patients in the early and late recovery phases following stroke," *Gait & Posture*, vol. 18, pp. 114–125, 2003.

[152] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Transactions on Multimedia*, vol. 15, pp. 1110–1120, 2013.

[153] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1737–1746, Austin, Texas, USA, 2012.

[154] P. J. Phillipsl, I. R. Sudecp Sarkari, P. Grotherl, and K. Bowyer, *The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm*, IEEE, Quebec City, Quebec, Canada, 2002.

[155] R. T. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pp. 366–371, Washington, DC, USA, 2002.

[156] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: multitarget detection and tracking," in *European Conference Computer Vision*, pp. 28–39, Prague, Czech Republic, 2004.

[157] V. D. Shet, V. S. N. Prasad, A. M. Elgammal, Y. Yacoob, and L. S. Davis, "Multi-cue exemplar-based nonparametric model for gesture recognition," in *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 656–662, Kolkata, India, 2004.

[158] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 2929–2936, Miami, FL, USA, 2009.

[159] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra, "Action categorization in soccer videos using string kernels," in *2009 Seventh International Workshop on Content-Based Multimedia Indexing*, pp. 13–18, Chania, Crete, Greece, 2009.

[160] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1290–1297, Providence, RI, USA, 2012.

[161] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *European Conference on Computer Vision*, pp. 392–405, Heraklion, Crete, Greece, 2010.

[162] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, pp. 971–981, 2013.

[163] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: a color-depth video database for human daily activity recognition," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 193–208, Springer, Barcelona, Spain, 2013.

[164] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *2012 IEEE International Conference on Robotics and Automation*, pp. 842–849, Saint Paul, MN, USA, 2012.

[165] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, "ConvNets-based action recognition from depth maps through virtual cameras and Pseudocoloring," in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1119–1122, Brisbane, Australia, 2015.

[166] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: a comprehensive multimodal human action database," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 53–60, Tampa, FL, USA, 2013.

[167] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, *Documentation Mocap Database hdm05*, Universit¨at Bonn, D-53117 Bonn, Germany, 2007.

[168] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 28–35, Providence, RI, USA, 2012.

[169] M. S. Ryoo and J. K. Aggarwal, "UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA)," in *IEEE International Conference on Pattern Recognition Workshops*, p. 4, Istanbul, Turkey, 2010.

[170] V. Bloom, D. Makris, and V. Argyriou, "G3d: a gaming action dataset and real time action recognition evaluation framework," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern*, pp. 7–12, Providence, RI, USA, 2012.

[171] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTU RGB+ D: a large scale dataset for 3D human activity analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019, Las Vegas, NV, USA, 2016.

[172] K. Soomro, A. R. Zamir, and M. Shah, *UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, arXiv Prepr*, Cornell University Library, Ithaca, NY, USA, 2012, http://arxiv.org/abs/1212.0402.

[173] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," in *Computer Vision - ECCV 2012*, pp. 872–885, Springer, Florence, Italy, 2012.

[174] A. Jalal, S. Kamal, and D. Kim, "Individual detection-tracking-recognition using depth activity images," in *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 450–455, Goyang, South Korea, 2015.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 4128183, 8 pages
https://doi.org/10.1155/2017/4128183

Research Article

An Interactive Care System Based on a Depth Image and EEG for Aged Patients with Dementia

Xin Dang, Bingbing Kang, Xuyang Liu, and Guangyu Cui

School of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China

Correspondence should be addressed to Xin Dang; xindang_tjpu@126.com

Received 24 February 2017; Accepted 14 May 2017; Published 18 July 2017

Academic Editor: Junfeng Gao

Due to the limitations of the body movement and functional decline of the aged with dementia, they can hardly make an efficient communication with nurses by language and gesture language like a normal person. In order to improve the efficiency in the healthcare communication, an intelligent interactive care system is proposed in this paper based on a multimodal deep neural network (DNN). The input vector of the DNN includes motion and mental features and was extracted from a depth image and electroencephalogram that were acquired by Kinect and OpenBCI, respectively. Experimental results show that the proposed algorithm simplified the process of the recognition and achieved 96.5% and 96.4%, respectively, for the shuffled dataset and 90.9% and 92.6%, respectively, for the continuous dataset in terms of accuracy and recall rate.

1. Introduction

The expected growth of the older adult population in China over the next 30 years will have an unprecedented impact on the healthcare system, especially in terms of supply and demand for healthcare workers. Moreover, the elders are always short of self-care ability and require manual care. Nurses have heavy working burden, especially when they are taking care of the high-risk and 24-hour guardian patients. In addition, the rehabilitation training always depends on the experienced therapist. In fact, the shortage of practitioners of nursing and rehabilitation is serious. Therefore, the development of cheap and efficient intelligent care equipment is a significant way to solve these problems.

With the development of automatic rollover, efficient bedsore care beds appeared firstly and got a social positive assessment in 2007. This research achieves some good results on the smart rollover and bedsore care facility aspects. However, due to the high cost and complexity in a hardware system and with the lack of automation, those techniques are still in the lab.

Another limitation of those healthcare facilities is the low efficiency of the communication between the equipment and elders. The help requirement of the elders always includes feeding, going to the toilet, rehabilitation and massage treatments, or chat with someone. However, there are some difficult states that are not well settled for traditional human-machine interface technologies, such as accent pronunciation, weak sound, and difficulty in moving. Thus, improving the efficiency and confidence of the elders in their interaction with healthcare facilities becomes one of the most important research topics, in favor of both self-care and rehabilitation and reducing the burden on their children.

With the increasing use of portable computing and communication devices, researchers try to use smartphones or tablet in improving the convenience in the healthcare interactions. However, because of the declining vision, dry fingers, complex operation process, and high-power consumption, these devices will be abandoned after a period of time. Furthermore, the development of somatosensory technology provides a superior interaction experience in the medical rehabilitation. Some low-cost somatosensory sensors provide a high practical interaction system by gesture recognition and speech recognition algorithm and are widely used in the active sports therapy and rehabilitation of patients [1–4].

For the neuropsychology rehabilitation service, Chang et al. developed a rehabilitation system, Kinect's Kinempt, based on the cognitive impairment [1]. This system utilized fast rehabilitation simulation according to the patient rehabilitation in a pizza shop catering. Task instructions in the form of real-time video get feedback on users. Two patients measured by excitation in the system voice performed pizza topping choices. If the action of the wrist extracted by Kinect matches the eating action, encourage voice will be provided; otherwise, an error message is displayed. The successful patient training rate of the rehabilitation treatment is between 20% and 60% for the absence of the intervention system. The success rate of patient training is over 90% with the rehabilitation treatment intervention. In order to slow down mental illness, for example, Alzheimer's disease (senile dementia), Chiang et al. developed an interactive game for elderly that can improve somatosensory cognitive function [5]. To complete the gaming experience by a plurality of operational tasks within the prescribed time, the system provides an evaluation procedure. The experiences using a Kinect video game are carried out for around 4 weeks. The reaction rate has improved as well as the hand-eye coordination.

For the physical rehabilitation service, Chang et al. developed a system based on the Kinect Kinerehab [6]. The athlete patients are under the rehabilitation guidance. The system uses Kinect motion image processing techniques to extract information about the patient. The patients are required to carry some rehabilitation operations such as consistently moving and lateral rising of the arms. The patient's joint position will be drawn, and Kinect computer database matching precisely calculates the degree of action in place. The results are fed back onto the display device. Meanwhile, patients are encouraged to test the effects of rehabilitation. In the absence of an intervention system stage, the accuracy rate of tested results is low; in a systematic intervention phase, the accuracy rate of measurement was increased significantly.

For the balance therapy rehabilitation service, Lange et al. combine the virtual reality and video game technology and developed 3D models for spinal cord injury and traumatic brain injury patients with a balance rehabilitation training game, which applied Unity3D engine into the development platform to develop [7–9]. During the game, the system reflects an image with virtual human posture and movement state of patients. A Kinect sensor tracks objects in real time. Patients could adjust the balance of the shoulder, elbow, and other parts through one received feedback information. To achieve real-time interaction with virtual people, they improve the balance of perception. Patients can perform flexion and external rotation, and offsite training in subjects such as passively assisting the shoulder in the medical division complete the scheduled action.

For the occupational therapy rehabilitation service, Gama et al. designed a shoulder and elbow exercise rehabilitation training system [10] based on Kinect. In the rehabilitation process, patients maintain a high shoulder and arm movement was held side bottom. If the measured results maintain the correct body posture, the system interface information bar will display a real-time activity of the joint. Else, if the action fails to meet the requirements, the system corrects the action by extracting a depth image captured by a Kinect sensor and calculating shoulder and elbow angles (angle of the two connection lines shoulder-elbow and shoulder-hip). Redress ways include correcting the curved portion of the elbow and plane deviated. When the angle between a normal vector of the shoulder-hand and a crown vector connecting the node plane is not equal to 90°, the system will show error and suggest that elbow movements of patients deviate from the crown plane; if the vector sum of the shoulder-elbow and elbow-hand is not equal to that of the shoulder-hand connection vector, the system will show error and point that a patient has a bent portion on the elbow.

Table 1: The label and features in the training set.

The types of requirement	Number of samples	Movement and mental features
Sleeping	192	Hand near the cheek and tired
Standing	198	Open arms and vibrant
Walking	202	Knee up and intent
Drinking	208	Hand near the mouth and thirsty
Eating	193	Hand near the mouth and hungry
Defecation	207	Head movement and defecating urgently
Urination	215	Head movement and urinating urgently
Calling the doctor	185	Hand movement and urgent
Nothing	200	None

For the interaction assistance in rehabilitation, Luo et al. use a Kinect sensor that recognizes a target after extracting patient gestures. The recognition result was converted into control commands. The commands are transferred to patients' wheelchairs through the internet to control the chairs intelligently, to help lower limb movement disorder patients in the rehabilitation life.

In this paper, we proposed a novel system based on the multimodal deep neural networks for the patient with dementia with special needs. The input features of the networks are extracted based on the depth image sensor (Kinect) and electroencephalogram sensor (OpenBCI). The output layer will result in a type recognition of the patient's help requirement.

This paper is organized as follows. Section 2 describes the collection of the training dataset, system flowchart, and the feature extraction. Section 3 gives the performance help requirement a recognition algorithm and then compares it with other similar methods. Finally, Section 5 gives a practical application of this system and future research directions.

2. Data and Methods

2.1. Dataset Collection. In the current investigation, 15 elderly patients aged 55–70 with limited mobility and vocality from Tianjin nursing home were used as study subjects. All the participants are not in good health conditions; 8 of

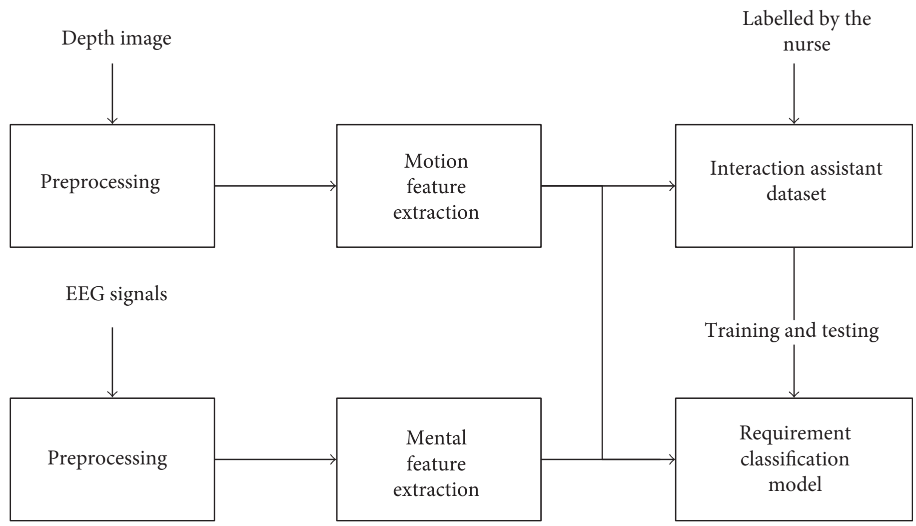


Figure 1: The system flowchart.

them are diagnosed as having dementia and 7 are diagnosed as having mild dementia. For each participant, 3-hour depth image and EEG data are recorded during their activity time with a continuous close-and-open eye movement as a synchronization pulse in the beginning of the data record. All the features are extracted from each event record and labeled with the types of the requirement by the nurses. After all, 1800 samples are packed into the dataset with 10 groups, for the cross-validation method in the neural network training and evaluation. The training set is shown in Table 1.

2.2. Overview of the System. The system acquired the record of the user's motion and mental states based on depth image and electroencephalogram signals by the Kinect and OpenBCI sensors, respectively. The motion features are extracted form a depth image after preprocessing with a filter, skeletal point tracking, region division, and normalization. The mental features are extracted from electroencephalogram signals after preprocessing with a bandpass filter and normalization. All the records of the interaction requirement activity are labelled by the nurse and collected into the dataset; the classification models are then trained and tested by the dataset. The flowchart of proposed system is shown in Figure 1.

2.3. The Feature Extraction of a Depth Image. The system analyzes the users' gesture by extracting 20 skeletal points from the depth image with a Kinect sensor. The three-dimensional coordinates of all the skeletal points are calculated and normalized using the spine skeletal coordinate as the origin [11]. The skeletal points are shown in Figure 2. Six regions are divided as the input feature of the system; they are cervical area, left wrist, right wrist, crotch, left knee, and right knee. Preprocess algorithms calculate the motion feature with the relative distance and direction of the regions, which were calculated by the location of the region center and their corresponding direction. Many activities in the dataset occur in 2-3 seconds; thus, in order to extract the feature changing in the help interaction, a three-segmentation scheme is applied for each record.

2.4. The Feature Extraction of EEG Signal. After the filtering process, then, this 8-dimensional data are put into the proposed system. For each channel, four band signals provided based on the frequency by the OpenBCI denoted δ (0.1–3 Hz), θ (4–7 Hz), α (8–12 Hz), and β (12–30 Hz). The electrode placement includes F3, F4, C3, C4, T5, T6, O1, and O2 as shown in Figure 3. Similar with those of the motion feature extraction, the mean and variance of three divisions of each activity record are calculated for the mental features.

3. Autoencoder and Classification Model

3.1. Autoencoder. The proposed algorithm consists of two parts: (1) sparse encoder—mainly used to study the character expression of the users' operation in healthcare, and (2) softmax category layer—recognizes the user's action through softmax based on the expression of sparse feature from the encoder in the hidden layer:

(1) Sparse autoencoder and unsupervised learning method

An autoencoder (AE) is a three-layer neural network that contains a visible layer, hidden layer, and reconstruction layer. Unsupervised learning and backpropagation (BP) algorithm and output vector equal to the input are used in the reconstructed layer.

A sparse autoencoder is a modified model as an encoder and obtains sparse characteristic through adding specific conditions to obtain the hidden layer in the training process [12–14]. We use the sigmoid function $f(a) = 1/(1 + \exp(-a))$ as the activation function. The feature extraction process of an AE includes two stages as shown in Figure 4. In the first stage, the user's motion and mental features x were mapped to the hidden layer z. The second phase generates and

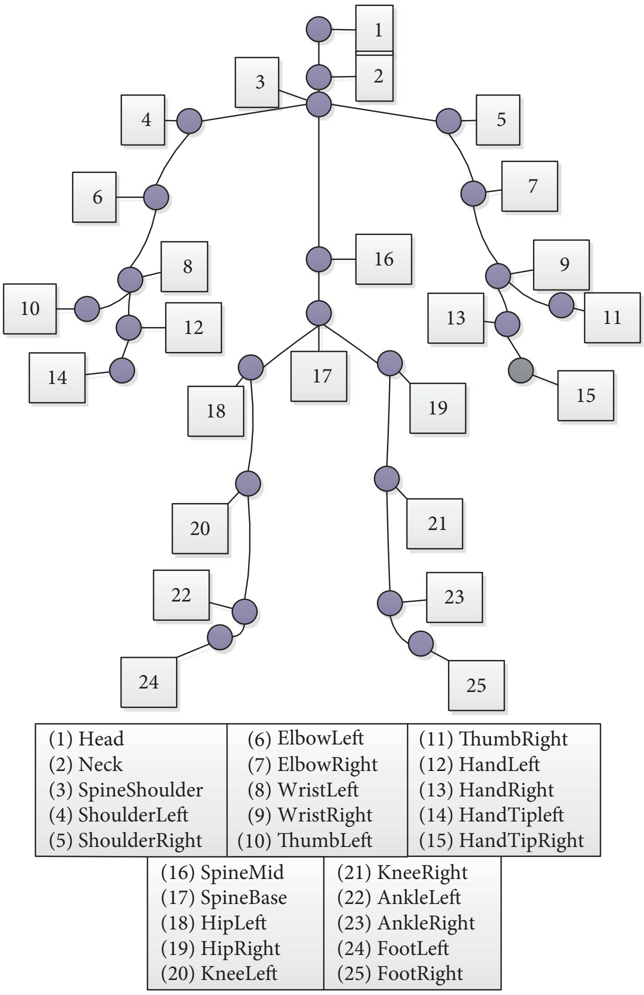


FIGURE 2: Skeleton tracking-based Kinect.

outputs y by decoding and reconstructing z. These two stages can be formulated as follows:

$$z = f(W_e n + b_e),$$
$$y = f(W_d n + b_d). \tag{1}$$

W_e and W_d are weight matrices of the encoder and the decoder, and b_e and b_d are the offset vectors. SAE training is used to minimize cost function $c(n, y)$ by adjusting the parameters (W_e, W_d, b_e, and b_d) using the BP algorithm.

$$\arg\min_{W_e, W_d, b_e, b_d} [c(n, y)], \tag{2}$$

where $c(n, y)$ denotes the error between the reconstructed layer and visible layer.

While input features are reconstructed in an output layer, high-level potential features related to users' requirement can be extracted in the hidden layer; the parameters of the hidden layer can be served as inherent characteristics of users' requirement. Hence, an n-dimensional input feature vector of the user's activity can be converted to h-dimensional potential features of users' requirement.

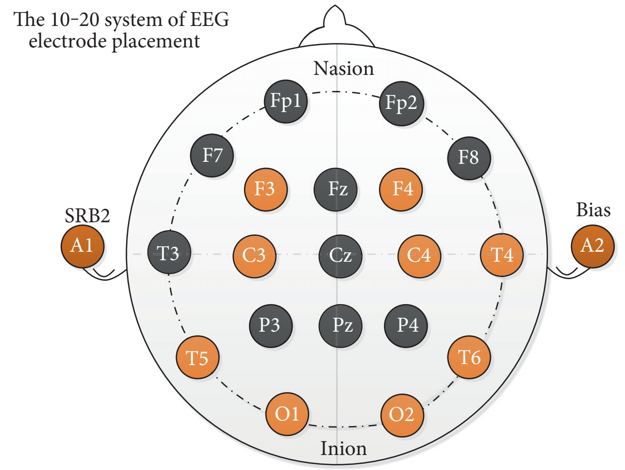


FIGURE 3: The electrode placement.

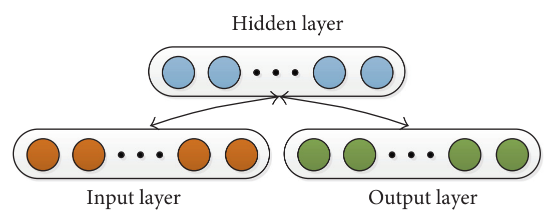


FIGURE 4: Schematic structure of an autoencoder.

The potential features $e(x)$ in the hidden layer can express input vectors in a compression method. When removing redundant relationship of input vectors after adding restrictions, $e(x)$ can explore the significant cross-correlation structure. This feature making the instance correspond to the input feature vector can be represented as a sparse mixture. Redundancy of joint distribution of any two features is compressed minimally in an SAE model. This model can ensure the robustness of the system when the input feature vectors have some damages. Therefore, this structure of the encoder achieves great success in the varieties of mode feature extraction. SAE cost function is defined as follows:

$$c(n, y) = \frac{1}{2q}\sum_{i=1}^{q}\left\| n^{(i)} - y^{(i)} \right\|_2^2 + \alpha \sum_{i=1}^{h} \mathrm{KL}(\rho \| \widehat{\rho}_i). \tag{3}$$

The first term is the average of differences among q input vectors, and the second is the sparse penalty term generated by h neurons in the hidden layer. The constant α denotes the control coefficient of sparse penalty. The KL divergence of two random variables in constraint entry can be formulated as follows:

$$\mathrm{KL}(\rho \| \widehat{\rho}_i) = \rho \log \frac{\rho}{\widehat{\rho}_i} + (1-\rho)\log\frac{1-\rho}{1-\widehat{\rho}_i}, \tag{4}$$

where $\widehat{\rho}_i = 1/q\sum_{j=1}^{q} zi(n^{(j)})$ is the activity of j neuron and ρ is the sparse parameter.

(2) Stacked autoencoder and supervised learning method

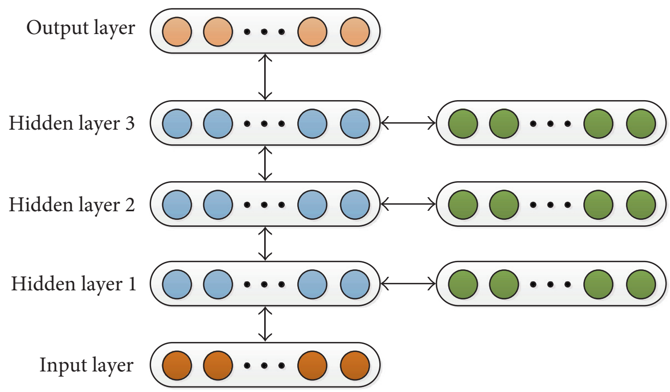


Figure 5: Stacked autoencoders.

A stacked autoencoder neural network is composed of multilayer sparse encoders. The output of the previous layer is the input of the subsequent layer. Characteristic expression of the last hidden layer will be put into a softmax classifier through the types of help requirement. The training process includes a pretraining stage and refine stage [2, 12, 15, 16]:

(a) Pretraining stage

A greedy scheme is applied in the training of SAE, the training method, as shown in Figure 5. The hidden layer of the ith autoencoder layer is the input of the next autoencoder.

(b) Refine stage

The proposed algorithm took the last hidden layer of the third autoencoder as the output layer of supervised training, by using the backpropagation algorithm. Each output unit has a corresponding label. Then, weights in the network are refined by the BP algorithm using the labelled data.

The softmax regression (SR) model used in the BP algorithm is shown in Figure 6. (z, L) is annotated training data in the model, where z is the expression of characteristic of the hidden layer and L is the classification label; users' activities are labeled from help requirement types from 1 to k.

The probability vector of the SR model $r_\theta(z)$ can be formulated as follows:

$$r_\theta(z) = \begin{bmatrix} p(l=1) \\ \dots \\ p(l=m) \end{bmatrix} = \frac{1}{\sum_{i=1}^{m} e^{\theta_i^t z}} \begin{bmatrix} e^{\theta_i^t z} \\ \dots \\ e^{\theta_m^t z} \end{bmatrix}, \tag{5}$$

where $\theta = [\theta_1^t, \dots, \theta_m^t]$. $p(l=m)$ is the predicted probabilities of the user's mth class label for each value of $m = 1, \dots, k$. The output of the SR model will be the help requirement class label with the highest probability results.

Thus, the cost function of the SR model can be formulated as follows:

$$J(\theta) = \frac{1}{q} \left[\sum_{i=1}^{q} \sum_{j=1}^{m} 1\{l=j\} \log \frac{e^{\theta_i^T z}}{\sum_{j=1}^{m} e^{\theta_i^T z}} \right], \tag{6}$$

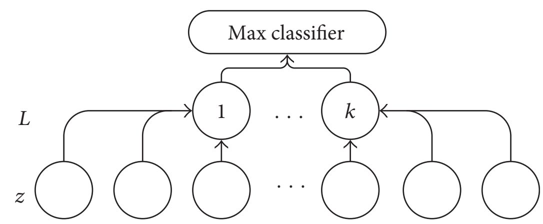


Figure 6: Softmax classification model.

where $l\{.\}$ is the indicator function, that is, $l\ \{true\} = 1$ and $l\{false\} = 0$. θ denotes all the parameters of our model and ensures that the cost function $J(\theta)$ is minimum.

(3) SAE-based multimodal classification

Due to the health condition, the aged always have an accent, weak sound, and limited body movement. For these situations, by imitating the workflow of a nurse in the healthcare, we extracted the patient's gestures and mental state from their body movement and EEG features. Furthermore, three types of the SAE-based multimodal models are proposed. And the details of them are shown as follows.

For the first type, the SAE model is pretrained and refined with the patient's motion and EEG features, respectively. The SAE network structures are shown in Figure 7.

For the second type, the input features of the motion and EEG are integrated at the first layer and pretrained with the integrated features in each hidden layer. The output layer is calculated by the integrated hidden layer. In the refine stage, the model is trained by the integrated features and labeled data. The network structure is shown in Figure 8.

For the third type, the input features of the motion and EEG are pretrained, respectively, at the second layer and integrated in the third layer. The output layer is calculated by the integrated hidden layer. In the refine stage, the model is trained by the two-modal features and labeled data. The network structure is shown in Figure 9.

Those three architectures are trying to imitate the workflow of the health nurses. The first model predicts the patient's requirement with the skeletal and EEG features, respectively; the second model predicts the patient's requirement with the integrated skeletal and EEG features with

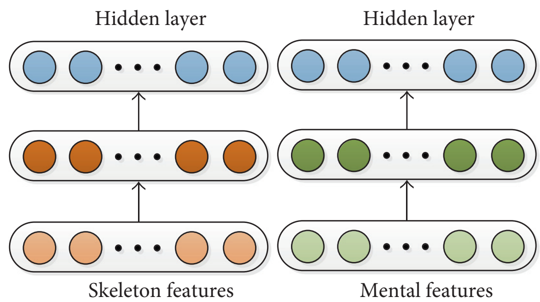


Figure 7: Single-modal classification model based on SAE.

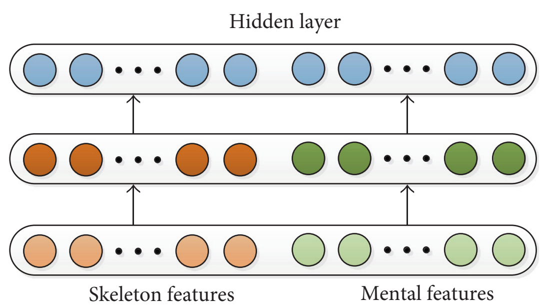


Figure 8: Multimodal classification model based on SAE.

limited use of the deep neural network as a high-level feature extraction; and the third model extracts a high-level feature in each hidden layer and integrates the high-level feature of the two-modal signals, then completes the classification by the SR model.

4. Experimental Results

SAE is built and trained based on the MATLAB Deep Learning Toolbox, which includes (1) SAE trained by the skeletal features, (2) SAE trained by EEG, and (3) SAE trained by the integrated data of both. All of the three SAEs contain three hidden layers.

For single-modal networks, the hidden layer size are set as 80 and the input sizes of the skeletal modal and EEG modal were set as 200 and 80, respectively. For a multimodal network, hidden layer size was set as 80; the input sizes of the skeletal modal and EEG modal were set as 200 and 80, respectively; and the size of the hidden layer was set as 160.

In the evaluation, the performance of a multiple recognition model based on SAE has been compared: classification based on skeletal features (skeleton), classification based on EEG feature (EEG), classification based on skeletal and EEG features (skeleton-EEG) (see Figure 6), and integration of two-modal features in hidden layers (integrated) (see Figure 7). Furthermore, as a comparison, the DTW method [17] is added into the evaluation.

In the evaluation, the accuracy P, the recall rate R, F1, and time consumption are used as the indicators; they are calculated as follows:

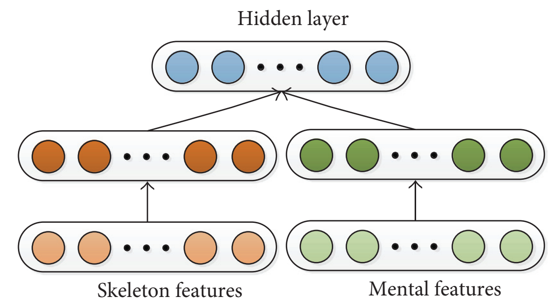


Figure 9: Multimodal classification model based on SAE.

$$P = \frac{A_g}{A_g + A_n} \times 100;$$
$$R = \frac{A_g}{A_g + A_n} \times 100; \quad (7)$$
$$\mathrm{F1} = \frac{2 \cdot P \cdot R}{P + R},$$

where A_g identified as the classification is distinguished correctly, A_n is misrecognition, and N_g means the recognition that the patient does nothing but was recognized with other classes.

The proposed methods are tested on two types of dataset:

(1) All classes of the help requirement record are extracted and shuffled into a dataset. The feature and label of each record are extracted and collected as the training data and test data sets.

(2) All raw data are divided into record samples by the frame. The frame length and frameshift are 3 seconds and 50%, respectively. The feature and label of each sample are extracted and collected as the training data and test data sets.

For the first type of dataset, the experiment results of the methods are shown in Table 2. Compared with that of three single-modal methods, the accuracy of the multimodal models achieved 5%–7% improvement under the same level of recall ratio. The integrated method showed the best accuracy and recall rate. The results of the skeleton-EEG method have a large improvement than those of the EEG or skeleton method in terms of accuracy and recall rate. The results of the skeleton method are little worse than those of the EEG-based model. It is suggested that the motion feature is insufficient for the activity recognition of the patient with dementia. However, skeleton features in the skeleton-EEG method and integrated method improve the performance in terms of both accuracy and recall rate. It is suggested that the multimodal method is better on exploring the characteristics of the physical and psychological activities in healthcare interaction for the dementia patients.

For the second type of dataset, the results are shown in Table 3. The performance is decreased in each term due to the frame division of the dataset. Compared to the traditional DWT algorithm, the SAE models achieved higher recognition

Table 2: Results of classification of the methods on the shuffled dataset.

Classifier	Accuracy (%)	Recall rate (%)	F1 measure	Time (s)
EEG	94.2	92.8	93.7	0.01
Skeleton	93.8	93.4	93.8	0.01
Skeleton-EEG	94.7	94.6	94.2	0.01
Integrated	96.5	96.4	96.2	0.01

Table 3: Results of classification of the methods on the continuous dataset.

Classifier	Accuracy (%)	Recall rate (%)	F1 measure	Time (s)
DTW [17]	84.2	90.0	89.6	0.035
EEG	90.1	87.6	87.7	0.01
Skeleton	89.3	86.4	86.9	0.01
Skeleton-EEG	90.7	90.4	89.5	0.01
Integrated	90.9	92.6	91.3	0.01

accuracy, despite that the single-modal models resulted in a little bit lower recall rate. Two multimodal models showed good performance both in the accuracy and in the recall rate. It is suggested that the multimodal model results in a significant improvement of the recall rate with the same level of the accuracy. The integrated model further increased the recall rate for the continuous input samples.

5. Conclusions

Healthcare and nursing robot achieve wide attention in recent years [18–21]. Somatosensory technology has been introduced into the activity recognition and healthcare interaction. However, due to the limited body movement of the patients, the traditional single-modal method is unsuitable to settle this interaction problem. On the other hand, with the development of the deep neural network technology, the depth model is applied as a high-level feature extractor and multimodal feature integrator. In order to develop an efficient and convenient interaction assistant system for nurses and patients with dementia, two novel multimodal SAE frameworks are proposed in this paper based on motion and mental features. The motion and mental features are extracted after the preprocessing of depth image and EEG signals acquired from Kinect and OpenBCI, respectively. The proposed algorithms simplify the acquisition and data processing under high action recognition ratio compared with the traditional DTW method and performed better in terms of both accuracy and recall rate.

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. 1403276 and the Research Program of Application Foundation and Advanced Technology of Tianjin under Grant nos. 14JCYBJC42300 and 14JCYBJC42400.

References

[1] Y. J. Chang, S. F. Chen, and A. F. Chuang, “A gesture recognition system to transition autonomously through vocational tasks for individuals with cognitive impairments,” *Research in Developmental Disabilities*, vol. 32, no. 6, pp. 2064–2068, 2011.

[2] D. González-Ortega, F. J. Díaz-Pernas, M. Martínez-Zarzuela, and M. Antón-Rodríguez, “A Kinect-based system for cognitive rehabilitation exercises monitoring,” *Computer Methods & Programs in Biomedicine*, vol. 113, no. 2, pp. 620–631, 2014.

[3] C. J. Su, C. Y. Chiang, and J. Y. Huang, “Kinect-enabled home-based rehabilitation system using dynamic time warping and fuzzy logic,” *Applied Soft Computing*, vol. 22, no. 5, pp. 652–666, 2014.

[4] C. Myers, L. Rabiner, and A. Rosenberg, “Performance tradeoffs in dynamic time warping algorithms for isolated word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 6, pp. 623–635, 1980.

[5] I. T. Chiang, J. C. Tsai, and S. T. Chen, “Using Xbox 360 Kinect games on enhancing visual performance skills on institutionalized older adults with wheelchairs,” in *2012 IEEE Fourth International Conference on Digital Game and Intelligent Toy Enhanced Learning*, pp. 263–267, Boston, 2012.

[6] Y. J. Chang, S. F. Chen, and J. D. Huang, “A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities,” *Research in Developmental Disabilities a Multidisciplinary Journal*, vol. 32, no. 6, pp. 2566–2570, 2011.

[7] B. Lange, S. Flynn, and A. Rizzo, “Initial usability assessment of off-the-shelf video game consoles for clinical game-based motor rehabilitation,” *Physical Therapy Reviews*, vol. 14, no. 9, pp. 355–363, 2009.

[8] B. Lange, C. Y. Chang, E. Suma, B. Newman, A. S. Rizzo, and M. Bolas, “Development and evaluation of low cost game-based balance rehabilitation tool using the Microsoft Kinect sensor,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1831–1834, Boston, 2011.

[9] B. Lange, S. Koenig, E. McConnell et al., “Interactive game-based rehabilitation using the Microsoft Kinect,” *2012 IEEE Virtual Reality Workshops (VRW)*, pp. 171-172, Orange County, 2012.

[10] A. D. Gama, T. Chaves, L. Figueiredo, and V. Teichrieb, “Poster: improving motor rehabilitation process through a natural interaction based system using Kinect sensor,” in *2012 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 145-146, Costa Mesa, 2012.

[11] R. Grosse, R. Raina, H. Kwon, and A. Ng, “Shift-invariance sparse coding for audio classification,” *Computer Science*, vol. 9, 2012.

[12] G. Hinton and R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, pp. 504–507, 2006.

[13] Q. V. Le, M. Ranzato, R. Monga et al., “Building high-level features using large scale unsupervised learning,” in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, 2012.

[14] H. Lee, A. Battle, R. Raina, and A. Ng, “Efficient sparse coding algorithms,” in *International Conference on Neural Information Processing Systems*, pp. 801–808, 2006, MIT Press.

[15] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Massachusetts, 2012.

[16] B. Moriarty, Z. Moy, and C. Amaba, “Utilizing explorable visual environments for experiential applications,” *Procedia Computer Science*, vol. 8, no. 8, pp. 261–266, 2012.

[17] S. Arteaga, J. Chevalier, A. Coile et al., “Low-cost accelerometry-based posture monitoring system for stroke survivors,” in *Proceedings of the 10th International ACM SIGACCESS*, pp. 243-244, 2008.

[18] D. Simonsen, J. Hansen, E. G. Spaich, and O. K. Andersen, “Kinect-Based Tele-rehabilitation System for Hand Function,” in *Replace, Repair, Restore, Relieve–Bridging Clinical and Engineering Solutions in Neurorehabilitation*, W. Jensen, O. Andersen and M. Akay, Eds., vol. 7 of Biosystems & Biorobotics, Springer, Cham, Berlin, 2014.

[19] P. Uttarwar and D. Mishra, “Development of a kinect-based physical rehabilitation system,” in *2015 Third International Conference on Image Information Processing (ICIIP)*, pp. 387–392, Himachal Pradesh, 2015.

[20] S. Li, P. N. Pathirana, and T. Caelli, “Multi-kinect skeleton fusion for physical rehabilitation monitoring,” in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5060–5063, Boston, 2014.

[21] M. Patanapanich, V. Vanijja, and P. Dajpratham, “Self-physical rehabilitation system using the microsoft kinect,” in *2014 International Conference on Information Technology Systems and Innovation (ICITSI)*, pp. 241–247, Bandung, 2014.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 2634389, 7 pages
https://doi.org/10.1155/2017/2634389

Research Article

An Evaluation of the Benefits of Simultaneous Acquisition on PET/MR Coregistration in Head/Neck Imaging

Serena Monti, Carlo Cavaliere, Mario Covello, Emanuele Nicolai, Marco Salvatore, and Marco Aiello

IRCCS SDN, Naples, Italy

Correspondence should be addressed to Serena Monti; smonti@sdn-napoli.it

Received 24 February 2017; Revised 2 May 2017; Accepted 16 May 2017; Published 18 July 2017

Academic Editor: Pan Lin

Coregistration of multimodal diagnostic images is crucial for qualitative and quantitative multiparametric analysis. While retrospective coregistration is computationally intense and could be inaccurate, hybrid PET/MR scanners allow acquiring implicitly coregistered images. Aim of this study is to assess the performance of state-of-the-art coregistration methods applied to PET and MR acquired as single modalities, comparing the results with the implicitly coregistration of a hybrid PET/MR, in complex anatomical regions such as head/neck (HN). A dataset consisting of PET/CT and PET/MR subsequently acquired in twenty-three patients was considered: performance of rigid (RR) and deformable (DR) registration obtained by a commercial software and an open-source registration package was evaluated. Registration accuracy was qualitatively assessed in terms of visual alignment of anatomical structures and qualitatively measured by the Dice scores computed on segmented tumors in PET and MRI. The resulting scores highlighted that hybrid PET/MR showed higher registration accuracy than retrospectively coregistered images, because of an overall misalignment after RR, unrealistic deformations and volume variations after DR. DR revealed superior performance compared to RR due to complex nonrigid movements of HN district. Moreover, simultaneous PET/MR offers unique datasets serving as ground truth for the improvement and validation of coregistration algorithms, if acquired with PET/CT.

1. Introduction

Integration of multimodal information carried out from different diagnostic imaging techniques is essential for a comprehensive characterization of the region under examination. Therefore, image coregistration has become crucial both for qualitative visual assessment [1] and for quantitative multiparametric analysis in research applications [2, 3] and clinical diagnosis, staging, and follow-up. Coregistration of complex data, such as diagnostic images, is typically computationally intense, and its result could also be inaccurate. This problem is intrinsically overcome by hybrid systems that allow acquiring simultaneously images that share the same coordinate system [4, 5].

In this field, the recently introduced integrated PET/MRI scanners represent the new frontier of molecular imaging. This new technology allows achieving in one-shot both functional information provided by positron emission tomography (PET) imaging and morpho-functional information with excellent soft tissue contrast provided by magnetic resonance imaging (MRI), increasing patient's compliance. The advantages of such a technology go beyond the mere combination of functional and morphological imaging: considering the wide range of MRI sequences and PET radiotracers available [6], the functional information of both MRI and PET may complement each another; moreover, due to the high spatial and contrast resolution of MRI, PET/MR imaging is becoming a straightforward clinical indication for local staging in complex anatomical regions such as head/neck [7], where it can help in delineating the tumor extent and lymph node involvement from the surrounding tissue [8–11]. Furthermore, PET/MRI can be useful for radiation therapy and presurgical treatment planning in head and neck cancer patients [12, 13].

With respect to separate acquisition of PET and MR, hybrid systems can certainly overcome the computational problem of the PET and MR coregistration, carrying out at same time PET and MR images of the same anatomical district that are, therefore, ideally coregistered.

Despite the undeniable advantages of hybrid solutions, their cost effectiveness is still far to be proven and the coregistration of multimodal information is frequently retrospectively obtained via software, combining images from a PET scanner with preexisting CT and MR, thus reducing the cost for new technology purchasing while offering renewed opportunities to advance PET, especially in underserved areas or under increasing economic constraints [14]. The problem of multimodal coregistration via software is commonly approached by algorithms consisting of an affine or rigid transformation followed by a free form deformation and using mutual information [15–17] as similarity measure. These algorithms are based, when available, on the coregistration of the anatomical information of CT component of PET/CT with MR, while the PET component can be transformed with the resulting deformation field, in order to guarantee more accurate coregistration of PET/MR data [13]. Retrospective coregistration via software has shown good performances also in the HN district [13, 18], but it is particularly challenging and technically demanding, mainly because of the varied patient positions used for the various scanners and the anatomic complexity of this region [10], which is subject to respiration, swallowing, and intrinsically nonrigid movements [19].

Looking at this scenario, our study is aimed to assess the performance of the state-of-the-art coregistration methods between PET and MR acquired as single modalities, comparing them with the intrinsic coregistration carried out by a hybrid PET/MR system, which is assumed to represent a ground truth for the assessment of retrospective coregistration. In particular, the performance of the state-of-the art rigid and deformable registration algorithms, implemented by a commercial software and an open-source registration package, was evaluated to appreciate their clinical suitability. Our work is based on a dataset of PET/MR and PET/CT of the HN district acquired during the same session, in order to exploit just a single administration of FDG-PET radiotracer.

2. Materials and Methods

The study was approved by the Institutional Review Board: 23 patients, with histologically confirmed HN malignancy (in early staging and in follow-up), were studied, after obtaining written informed consent. Table 1 shows clinical details for each patient.

2.1. Imaging Protocol. All subjects underwent a single-injection dual imaging protocol including PET/CT and subsequent PET/MR, so that no additional injection was required and any additional radiation exposure for the patients was avoided. The examination protocol consisted of the following steps: patients fasted for at least 6 h before scanning; just before the injection, the blood glucose level was measured in order to ensure a value below 150 mg/dL; and the patients were injected with about 400 MBq of

Table 1: Patient's cohort examined during this study. For each patient, in addition to personal details, the site of malignancy is specified.

ID	Age	Sex	Site
pt1	60	M	Rhinopharynx
pt2	68	M	Oropharynx
pt3	61	M	Tongue
pt4	70	M	Larynx
pt5	52	F	Hypopharynx
pt6	61	M	Larynx
pt7	56	M	Tongue
pt8	35	F	Larynx (neg)
pt9	72	M	Larynx
pt10	65	M	Rhinopharynx
pt11	51	M	Larynx (neg)
pt12	70	M	Larynx
pt13	53	M	Rhinopharynx
pt14	43	M	Oropharynx
pt15	68	M	Larynx
pt16	68	M	Larynx
pt17	83	M	Tongue
pt18	68	M	Skull base
pt19	86	F	Thyroid
pt20	33	M	Laterocervical
pt21	70	M	Larynx
pt22	58	F	Larynx
pt23	43	M	Larynx (neg)

[18F]-FDG, depending on their body weight. After an uptake period of 80 minutes, patients underwent PET/CT scanning and, soon after PET/CT, they underwent PET/MR examination.

2.1.1. PET/CT Acquisition. PET/CT acquisition was performed on a Gemini TF (Philips Medical Systems, Best, The Netherlands). PET data was acquired in sinogram mode for 15 minutes with a matrix size of 144×144. A 3-dimensional attenuation-weighted ordered-subsets expectation maximization iterative reconstruction algorithm (AW OSEM 3D) was applied with 3 iterations and 21 subsets, Gaussian smoothing of 4 mm in FWHM, and a zoom of 1. The CT consisted of a low-dose scan (120 kV, 80 mA). Patient position was supine with his arms resting at the side.

2.1.2. PET/MR Acquisition. PET/MR was performed on a Biograph mMR (Siemens Healthcare, Erlangen, Germany). Bed position was established in order to get a full coverage of the head/neck region. Also, these PET data were reconstructed with an AW OSEM 3D iterative reconstruction algorithm applied with 3 iterations and 21 subsets, Gaussian smoothing of 4 mm in full width at half maximum, and a zoom of 1. MR attenuation correction was performed via a segmentation approach based on 2-point Dixon MRI sequences. The MRI protocol was performed with dedicated head and neck coils. The MRI sequences taken into account for this study were T2-weighted short time inversion recovery (STIR) acquired in coronal direction (TR/TE/TI = 5000/84/220, one acquired signal, voxel size = 0.4 × 0.4 × 3.5 mm).

2.2. Data Processing. Image registration strategies were developed based on CT and MR data only, with the MR coronal STIR acquisition serving as fixed image and the CT as moving, whereas the PET component from PET/CT data was transformed only later by the resulting deformation field into the same coordinate system of PET/MR. The entire registration process was performed twice, using two different tools: the freely available, open-source registration package Elastix (http://elastix.isi.uu.nl) [20] and the tool for deformable image registration included in the commercial software XD3 (Mirada Medical Ltd., Oxford, United Kingdom) [21]. In the following, we will refer to the set composed by PET and MR acquired on the hybrid PET/MR scanner as PETMRo and to the set composed by PET from PET/CT retrospectively coregistered to MR from PET/MR as PETMRreg. Moreover, registration performed by means of Elastix or Mirada will be superscripted with ELX or MRD, respectively, while the suffixes RR and DR will indicate rigid and deformable registration, respectively.

2.2.1. Image Registration with Elastix. Elastix is a command line-driven program based on the Insight Toolkit (ITK) registration framework (open source: National Library of Medicine, www.itk.org). The registration parameters were selected based on previous work on multimodal deformable image registration for integration of PET/MR into radiotherapy treatment planning for head and neck [13]. First, a rigid registration (RR) was performed and the resulting transform was used as a starting point for deformable registration (DR), by means of B-spline transform [15]. Both RR and DR were performed with a three-level multiresolution approach, using Gaussian smoothing (sigma = 8.0, 4.0, and 4.0 in x and y direction and sigma = 2.0, 1.0, and 0.5 in z direction to take into account voxel anisotropy) without downsampling. A localized version of mutual information (LMI) was considered as similarity measure (Mattes Mutual Information) [22], and a stochastic gradient descent optimizer [23] was chosen to minimize it. In detail, for RR, LMI metric was computed with 64 bins, 2000 samples, and a maximum of 500 iterations for each resolution. For DR, a bending energy penalty (BEP) term was calculated [15] to regularize the transformation; the metric (sum of LMI and BEP) was computed with 60 bins, 10,000 samples, and a maximum of 5000 iterations for each resolution. After the deformation field that maps the CT into the space coordinates of MR was computed, the PET from PET/CT was accordingly warped using Transformix, another command line-driven program based on ITK that applies a known transformation to an input image.

2.2.2. Image Registration with XD3. XD3 is a Mirada's commercial platform that provides a full suite of practical applications for multimodal image viewing, including rigid and deformable registration. After a first step of RR between MR and CT, a DR was performed using Mirada's multimodal deformable image registration algorithm that optimizes a proprietary form of a mutual information-based similarity function [15, 16, 24] over a radial basis function (RBF) transformation model. Default parameters for MR-CT unsupervised registration were used. Finally, the PET from PET/CT was automatically warped, once the transformation that registers CT to MR was computed.

Table 2: Scoring system used to evaluate the registration quality of PET with MR images.

Score	Meaning
0	Case unusable
1	Alignment of major anatomical structures is sufficient but localization of tumors is not perfectly corresponding
2	Alignment of major anatomical structures is good and localization of tumors corresponds in PET and MRI

2.3. Image Evaluation. Registration accuracy was qualitatively and qualitatively evaluated in five sets of images for each patient: PETMRo, $\mathrm{PETMRreg_{ELX}}^{\mathrm{RR}}$, $\mathrm{PETMRreg_{ELX}}^{\mathrm{DR}}$, $\mathrm{PETMRreg_{MRD}}^{\mathrm{RR}}$, and $\mathrm{PETMRreg_{MRD}}^{\mathrm{DR}}$. Qualitative evaluation was performed by two clinical reviewers: one is a nuclear medicine physician who is also licentiate in diagnostic radiology and the other is a radiologist who is also licentiate in nuclear medicine. Images were analyzed in the coronal plane on the freely available medical imaging platform medInria [25], which allows visualization and fusion of both NIfTI files and DICOM files. The observers reviewed the five image set for each patient evaluating the alignment of the major anatomical structures. Then they identified, in PET and MR images, the localization and the extent of the primary tumor and metastasis to regional lymph nodes, and, on these bases, they independently rated the registration quality of each tested method using the scoring system defined in Table 2. Neither reader was aware of the results of other imaging studies, histopathologic findings, or clinical data.

In order to obtain also a quantitative evaluation of the registration accuracy, the two clinicians were asked to segment the primary tumor of each patient using the freely available software ITK-SNAP (http://www.itksnap.org) [26]. The radiologist manually contoured the lesion in the T2-weighted coronal image. The nuclear medicine physician used the user-guided 3D active contour segmentation implemented in ITK-SNAP to semiautomatically segment the primary tumor, after having initialized the process with the placement of a spherical seed. For each patient, he repeated five times the operation (once for each PET of the five sets). The obtained segmentations were then used to compute the Dice score (Dice, 1945 number 39) between MR and PET from each set.

2.4. Statistical Analysis. A Friedman statistics with successive multicomparative analysis were used to evaluate the statistical differences in visual ratings between implicitly coregistered PET/MR and the results of coregistration software. Further comparisons between the single steps of the different

Table 3: Evaluation results: registration accuracy scores expressed as mean ± standard deviation.

	PETMRo	$\text{PETMRreg}_{\text{ELX}}{}^{\text{RR}}$	$\text{PETMRreg}_{\text{ELX}}{}^{\text{DR}}$	$\text{PETMRreg}_{\text{MRD}}{}^{\text{RR}}$	$\text{PETMRreg}_{\text{MRD}}{}^{\text{DR}}$
Qualitative score	2.0 ± 0.0	0.70 ± 0.84	1.35 ± 0.70	0.76 ± 0.92	0.89 ± 0.80
Dice score	0.82 ± 0.07	0.36 ± 0.23	0.45 ± 0.24	0.37 ± 0.22	0.41 ± 0.23

coregistration software were evaluated by means of a Wilcoxon's signed-rank statistic test, as done in [27]. Similarly, statistical differences in Dice scores were tested with ANOVA and paired Student's t-test. Statistical analysis was performed using Matlab (MATLAB R2014b, Math-Works, Natick, MA). Differences at $p < 0.05$ were considered to be statistically significant.

3. Results

Table 3 shows the mean qualitative scores over the two observers and the Dice score for each patient and for each set considered.

Comparing PETMRo with registrations performed by Elastix and Mirada, their differences resulted to be statistically significant both at qualitative (Friedman test p values: $\text{PETMRo/PETMRreg}_{\text{ELX}}{}^{\text{RR}}/\text{PETMRreg}_{\text{ELX}}{}^{\text{DR}} = 2.92 \cdot 10^{-7}$, $\text{PETMRo/PETMRreg}_{\text{MRD}}{}^{\text{RR}}/\text{PETMRreg}_{\text{MRD}}{}^{\text{DR}} = 4.56 \cdot 10^{-7}$) and quantitative analysis (ANOVA p values: $\text{PETMRo/PETMRreg}_{\text{ELX}}{}^{\text{RR}}/\text{PETMRreg}_{\text{ELX}}{}^{\text{DR}} = 4.83 \cdot 10^{-11}$, $\text{PETMRo/PETMRreg}_{\text{MRD}}{}^{\text{RR}}/\text{PETMRreg}_{\text{MRD}}{}^{\text{DR}} = 3.80 \cdot 10^{-12}$). For each patient, the scores highlighted that PETMRo set showed a higher (at the most equal) registration accuracy than the other fused sets of images.

This superiority was statistically significant in all the comparison for qualitative (Wilcoxon test p values: $\text{PETMRo/PETMRreg}_{\text{ELX}}{}^{\text{RR}} = 1.25 \cdot 10^{-7}$, $\text{PETMRo/PETMRreg}_{\text{ELX}}{}^{\text{DR}} = 0.02$, $\text{PETMRo/PETMRreg}_{\text{MRD}}{}^{\text{RR}} = 2.46 \cdot 10^{-6}$, $\text{PETMRo/PETMRreg}_{\text{MRD}}{}^{\text{DR}} = 3.50 \cdot 10^{-5}$) and quantitative scores (t-test p values: $\text{PETMRo/PETMRreg}_{\text{ELX}}{}^{\text{RR}} = 1.57 \cdot 10^{-10}$, $\text{PETMRo/PETMRreg}_{\text{ELX}}{}^{\text{DR}} = 2.22 \cdot 10^{-8}$, $\text{PETMRo/PETMRreg}_{\text{MRD}}{}^{\text{RR}} = 1.45 \cdot 10^{-10}$, $\text{PETMRo/PETMRreg}_{\text{MRD}}{}^{\text{DR}} = 1.20 \cdot 10^{-9}$).

The registration results of PET with MR images after a RR step showed an overall misalignment due to different patient positioning, both for Elastix and Mirada results, with differences between the two methods that were not statistically significant both at the qualitative (Wilcoxon test p value $\text{PETMRreg}_{\text{ELX}}{}^{\text{RR}}/\text{PETMRreg}_{\text{MRD}}{}^{\text{RR}} = 0.73$) and at the quantitative scores (t-test p value $\text{PETMRreg}_{\text{ELX}}{}^{\text{RR}}/\text{PETMRreg}_{\text{MRD}}{}^{\text{RR}} = 0.75$).

If a DR step was performed, a significant improvement could be obtained, but also unrealistic deformations or moderate and smooth volume expansions and compressions could occur, leading to a good or sufficient alignment of major anatomical structures but local misregistration of tumors. However, looking at the scores of the single patients, after a DR step, the accuracy of registration tended to improve for both Elastix and Mirada results. This improvement was statistically significant for registration performed with Elastix (Wilcoxon test p value $\text{PETMRreg}_{\text{ELX}}{}^{\text{RR}}/\text{PETMRreg}_{\text{ELX}}{}^{\text{DR}} = 0.01$, t-test p value $\text{PETMRreg}_{\text{ELX}}{}^{\text{RR}}/\text{PETMRreg}_{\text{ELX}}{}^{\text{DR}} = 1.4 \cdot 10^{-3}$) and not statistically significant for Mirada (p value $\text{PETMRreg}_{\text{MRD}}{}^{\text{RR}}/\text{PETMRreg}_{\text{MRD}}{}^{\text{DR}} = 0.85$, p value $\text{PETMRreg}_{\text{MRD}}{}^{\text{RR}}/\text{PETMRreg}_{\text{MRD}}{}^{\text{DR}} = 0.09$)

Comparing the results obtained from the two registration tools, the scores of the DR output with Elastix were generally higher than those of the DR outputs, but the arisen differences between these two sets were statistically significant only at the qualitative assessment (Wilcoxon test p value $\text{PETMRreg}_{\text{ELX}}{}^{\text{DR}}/\text{PETMRreg}_{\text{MRD}}{}^{\text{DR}} = 0.02$, t-test p value $\text{PETMRreg}_{\text{ELX}}{}^{\text{DR}}/\text{PETMRreg}_{\text{MRD}}{}^{\text{DR}} = 0.32$.

In conclusion, while at the quantitative assessment, hybrid PET/MR definitely outperforms retrospective registration; at the qualitative score in the 25% of the cases, all retrospective coregistration methods showed results that were comparable with PETMRo in terms of alignment of major anatomical structures and tumors (Figure 1). In the remaining 75%, PETMRo exhibited an overall superiority. In detail, the 17% of these cases showed a slightly better performance of Elastix-based registration, in particular the DR step, in comparison with Mirada; one case showed a better performance of Mirada; in the remaining ones, problems of misalignment in RR steps and/or volume variations for DR steps were visible (Figure 2).

4. Discussion

In this work, four different strategies for the coregistration of PET and MR in the HN region were qualitatively and quantitatively evaluated, with the purpose of comparing them with the intrinsic coregistration of simultaneous PET/MR, which is assumed to represent a ground truth for the assessment of retrospective coregistration.

To our knowledge, this is the first reported study to have investigated the validity of retrospectively coregistered PET/MR of HN district using images obtained from different modalities in terms of localization and extent of the primary tumor and metastasis to regional lymph nodes and to have compared the accuracy of anatomical structure alignment and tumor localization with the intrinsic coregistered simultaneous PET/MR.

Kanda et al. [9] assessed the clinical value of retrospective image coregistration of neck MRI and [18F]-FDG PET for loco-regional extension and nodal staging of neck cancer in 30 patients, comparing it with PET/CT fusion. Although they used manual registration, they hypothesized that simultaneous PET/MR technology would minimize the drawbacks of retrospective PET/MR coregistration strategy, such as local misregistration, generating better-quality fusion images, as can be confirmed by our study.

The same has been studied by Loeffelbein et al. [11] that compared their retrospective coregistration results obtained

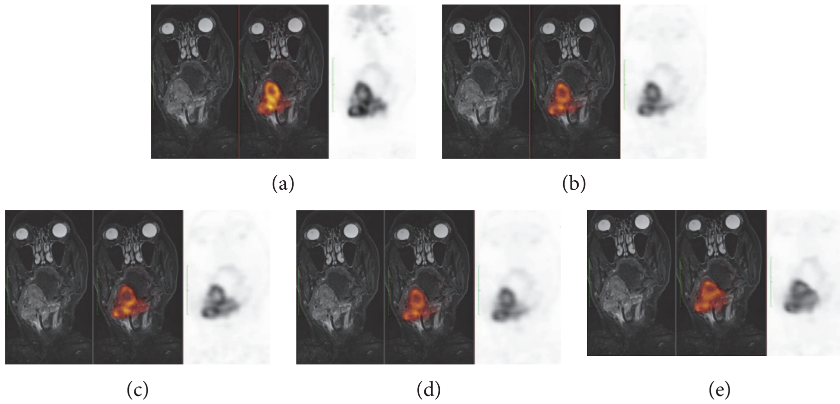


Figure 1: Example of qualitative well ranked coregistration results. From left to right: coronal MR image, fused PET/MR, and PET image from (a) PETMRo, (b) $\mathrm{PETMRreg_{ELX}}^{RR}$, (c) $\mathrm{PETMRreg_{ELX}}^{DR}$, (d) $\mathrm{PETMRreg_{MRD}}^{RR}$, and (e) $\mathrm{PETMRreg_{MRD}}^{DR}$. Both RR and DR with Elastix, (b) and (c), respectively, and Mirada, (d) and (e), respectively, show results comparable with intrinsic coregistration of simultaneous PET/MR (a). The Dice scores for this case are PETMRo = 0.95, $\mathrm{PETMRreg_{ELX}}^{RR} = 0.85$, $\mathrm{PETMRreg_{ELX}}^{DR} = 0.86$, $\mathrm{PETMRreg_{MRD}}^{RR} = 0.89$, and $\mathrm{PETMRreg_{MRD}}^{DR} = 0.90$.

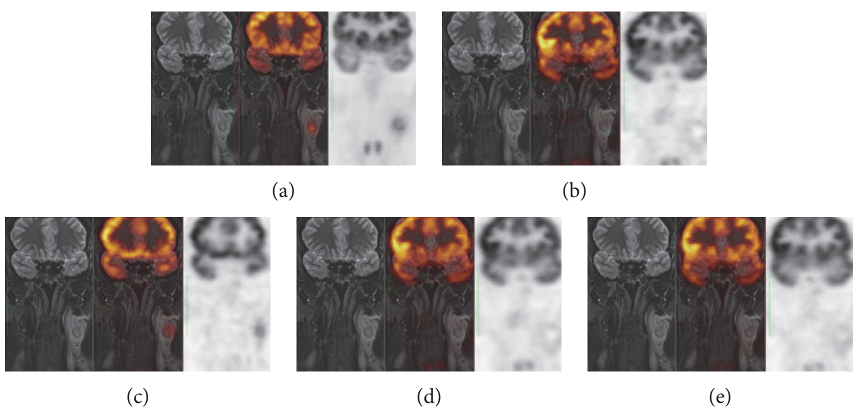


Figure 2: Example of poorly ranked coregistration results. From left to right: coronal MR image, fused PET/MR, and PET image from (a) PETMRo, (b) $\mathrm{PETMRreg_{ELX}}^{RR}$, (c) $\mathrm{PETMRreg_{ELX}}^{DR}$, (d) $\mathrm{PETMRreg_{MRD}}^{RR}$, and (e) $\mathrm{PETMRreg_{MRD}}^{DR}$. RR with Elastix (b) and Mirada (d) shows an overall misalignment of the brain contour between PET and MR. This misalignment is only partially recovered by DR with Mirada (e) and better recovered by DR with Elastix (c). However, the lymph node tumor is completely absent in the PET component of $\mathrm{PETMRreg_{MRD}}^{RR}$ and its localization is not perfectly corresponding in $\mathrm{PETMRreg_{ELX}}^{DR}$ (c) as in PETMRo (a). The Dice scores for this case are PETMRo = 0.79, $\mathrm{PETMRreg_{ELX}}^{RR} = 0.26$, $\mathrm{PETMRreg_{ELX}}^{DR} = 0.52$, $\mathrm{PETMRreg_{MRD}}^{RR} = 0.19$, and $\mathrm{PETMRreg_{MRD}}^{DR} = 0.25$.

by a commercial software with side-by-side analysis of single modality PET and MRI in a group of thirty patients.

In neither of these two studies, the authors had data sets coming from simultaneous PET/MR available to use as gold standard for the evaluation of registration performances.

Leibfarth et al. [13] developed an accurate and robust registration strategy on a dataset of eight patients consisting of an FDG PET/CT and a subsequently acquired PET/MR of HN with the aim of integrating combined PET/MR data into RT treatment planning. We started from this work for the implementation of registration with Elastix, but we took advantages of a wider dataset and we also evaluate registration performed by a commercial software optimized for a clinical workflow.

Our results showed that comparison of rigid versus deformable registration revealed superior performance for deformable registration both for Elastix and Mirada. This is due to the complex movements of this region, which are intrinsically nonrigid and hence cannot be completely recovered by a rigid transformation with only six degrees of freedom. With regard to deformable registration, although Elastix showed better performance than Mirada at least in the qualitative evaluation, it was computationally more intense. The tested software used different similarity measures and different transform basis: although Mirada uses RBF transformation model, which is more accurate and faster than the B-Spline used in this work for Elastix registration, it is completely embedded. Consequently, internal registration parameters, such as deformation field smoothness, degrees of freedom, and similarity function sensitivity, are automatically tuned by the software on the basis of the considered modalities. On the other side, the registration scheme and parameters used for DR in Elastix are the result of a previous optimization study [13] and are designed for the specific application in MR-CT registration of HN district; in particular, B-spline parametrization in conjunction with BEP is chosen to favor a smooth and reasonable transform. LMI is advantageous in the case of spatial intensity distortions and for multimodal registration if one intensity class corresponds to a specific tissue type in one imaging modality and to different tissue types in the other imaging modality [13]. Moreover, as expected, coregistered PET/MR images from hybrid scanner carried out the best performances, since they are inherently free from the problems of misalignment, local misregistration, and unrealistic deformation field.

We believe that performances of software-based coregistration method in districts subject to nonrigid movements, such as HN, could be undoubtedly improved by means of support structures, as head masks, designed for immobilizing the patient during the acquisitions. In addition, registration algorithms could benefit from users' supervision for preliminary manual step, in order to start from an optimal rigid alignment that could improve the performances of successive automatic deformable steps, making them more feasible in terms of computational time. Both these issues are out-of-the-scopes of this work that, although it is aimed to evaluate the clinical suitability of retrospective coregistration in comparison to intrinsic coregistration of simultaneous PET/MR, limits the investigation to a fully automated perspective.

In conclusion, our findings show that, regarding the complex case of PET/MR of HN district, there is a wide room for improvement of software-based coregistration algorithms, since, at present, they are definitely outperformed by the intrinsic coregistration of simultaneous PET/MR that overcomes the above-named problem of retrospective coregistration, as hypothesized in previous works [9, 10]. In this direction, simultaneous PET/MR imaging, which hence offers unique datasets when acquired together with PET/CT during the same session, could also serve as ground truth for the validation of improved coregistration algorithms.

Conflicts of Interest

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

Acknowledgments

This work was partially supported by the RRC-2015-2360454 of the Italian Minister of Health and by the Italian project PON03PE_00128_1 "eHealthNet: Software ecosystem for Electronic Health".

References

[1] M. Aiello, S. Monti, M. Inglese et al., "A multi-modal fusion scheme for the enhancement of PET/MR viewing," *EJNMMI Physics*, vol. 2, 2015, Springer.

[2] S. Monti, S. Cocozza, P. Borrelli et al., "MAVEN: an algorithm for multi-parametric automated segmentation of brain veins from gradient echo acquisitions," *IEEE Transactions on Medical Imaging*, vol. 36, no. 5, pp. 1054–1065, 2017.

[3] S. Monti, G. Palma, P. Borrelli et al., "A multiparametric and multiscale approach to automated segmentation of brain veins," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3041–3044, Milan, 2015.

[4] H. Zaidi and A. D. Guerra, "An outlook on future design of hybrid PET/MRI systems," *Medical Physics*, vol. 38, no. 10, pp. 5667–5689, 2011.

[5] G. Delso, S. Fürst, B. Jakoby et al., "Performance measurements of the Siemens mMR integrated whole-body PET/MR scanner," *Journal of Nuclear Medicine*, vol. 52, no. 12, pp. 1914–1922, 2011.

[6] G. Antoch and A. Bockisch, "Combined PET/MRI: a new dimension in whole-body oncology imaging?" *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 36, Supplement 1, pp. S113–S120, 2009.

[7] P. Veit-Haibach, F. P. Kuhn, F. Wiesinger, G. Delso, and G. von Schulthess, "PET–MR imaging using a tri-modality PET/CT–MR system with a dedicated shuttle in clinical routine," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 26, no. 1, pp. 25–35, 2013.

[8] C. Buchbender, T. A. Heusner, T. C. Lauenstein, A. Bockisch, and G. Antoch, "Oncologic PET/MRI, part 1: tumors of the brain, head and neck, chest, abdomen, and pelvis," *Journal of Nuclear Medicine*, vol. 53, no. 6, pp. 928–938, 2012.

[9] T. Kanda, K. Kitajima, Y. Suenaga et al., "Value of retrospective image fusion of 18F-FDG PET and MRI for preoperative staging of head and neck cancer: comparison with PET/CT and contrast-enhanced neck MRI," *European Journal of Radiology*, vol. 82, no. 11, pp. 2005–2010, 2013.

[10] D. J. Loeffelbein, M. Souvatzoglou, V. Wankerl et al., "PET-MRI fusion in head-and-neck oncology: current status and implications for hybrid PET/MRI," *Journal of Oral and Maxillofacial Surgery*, vol. 70, no. 2, pp. 473–483, 2012.

[11] D. Loeffelbein, M. Souvatzoglou, V. Wankerl et al., "Diagnostic value of retrospective PET-MRI fusion in head-and-neck cancer," *BMC Cancer*, vol. 14, no. 1, p. 846, 2014.

[12] S. Partovi, A. Kohan, C. Rubbert et al., "Clinical oncologic applications of PET/MRI: a new horizon," *American Journal of Nuclear Medicine and Molecular Imaging*, vol. 4, no. 2, p. 202, 2014.

[13] S. Leibfarth, D. Mönnich, S. Welz et al., "A strategy for multimodal deformable image registration to integrate PET/MR into radiotherapy treatment planning," *Acta Oncologica*, vol. 52, no. 7, pp. 1353–1359, 2013.

[14] R. L. Bridges, "Software fusion: an option never fully explored," *Journal of Nuclear Medicine*, vol. 50, no. 5, pp. 834–836, 2009.

[15] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[16] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[17] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "PET-CT image registration in the chest using free-form deformations," *IEEE Transactions on Medical Imaging*, vol. 22, no. 1, pp. 120–128, 2003.

[18] V. Fortunati, R. F. Verhaart, F. Angeloni et al., "Feasibility of multimodal deformable registration for head and neck tumor treatment planning," *International Journal of Radiation Oncology, Biology, and Physics*, vol. 90, no. 1, pp. 85–93, 2014.

[19] M. Becker and H. Zaidi, "Imaging in head and neck squamous cell carcinoma: the potential role of PET/MRI," *The British Journal of Radiology*, vol. 87, no. 1036, 2014.

[20] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[21] M. J. Gooding, C. L. Eccles, M. Fuss et al., "Assessing the quality of deformable CT-MR registration for the purpose of multimodal radiotherapy contouring," *International Journal of Radiation Oncology, Biology, and Physics*, vol. 81, no. 2, pp. S812–S813, 2011.

[22] S. Klein, U. A. van der Heide, I. M. Lips, M. van Vulpen, M. Staring, and J. P. Pluim, "Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information," *Medical Physics*, vol. 35, no. 4, pp. 1407–1417, 2008.

[23] S. Klein, J. P. Pluim, M. Staring, and M. A. Viergever, "Adaptive stochastic gradient descent optimisation for image registration," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 227–239, 2009.

[24] W. M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35–51, 1996.

[25] N. Toussaint, J.-C. Souplet, and P. Fillard, "MedINRIA: medical image navigation and research tool by INRIA," in *Proceedings of MICCAI Workshop on Interaction in Medical Image Analysis and Visualization*, 2007.

[26] P. A. Yushkevich, J. Piven, H. C. Hazlett et al., "User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[27] L. Pace, E. Nicolai, A. Luongo et al., "Comparison of whole-body PET/CT and PET/MRI in breast cancer patients: lesion detection and quantitation of 18F-deoxyglucose uptake in lesions and in normal organ tissues," *European Journal of Radiology*, vol. 83, no. 2, pp. 289–296, 2014.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 9296354, 12 pages
https://doi.org/10.1155/2017/9296354

Research Article

Nonrigid Registration of Prostate Diffusion-Weighted MRI

Lei Hao,[1,2] Yali Huang,[1] Yuehua Gao,[1] Xiaoxi Chen,[3] and Peiguang Wang[1]

[1]*College of Electronic Information Engineering, Hebei University, Baoding 071000, China*
[2]*Key Laboratory of Digital Medical Engineering of Hebei Province, College of Electronic and Information Engineering, Hebei University, Baoding 071000, China*
[3]*Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200127, China*

Correspondence should be addressed to Peiguang Wang; pgwang@hbu.edu.cn

Received 21 February 2017; Accepted 23 April 2017; Published 27 June 2017

Academic Editor: Pan Lin

Motion and deformation are common in prostate diffusion-weighted magnetic resonance imaging (DWI) during acquisition. These misalignments lead to errors in estimating an apparent diffusion coefficient (ADC) map fitted with DWI. To address this problem, we propose an image registration algorithm to align the prostate DWI and improve ADC map. First, we apply affine transformation to DWI to correct intraslice motions. Then, nonrigid registration based on free-form deformation (FFD) is used to compensate for intraimage deformations. To evaluate the influence of the proposed algorithm on ADC values, we perform statistical experiments in three schemes: no processing of the DWI, with the affine transform approach, and with FFD. The experimental results show that our proposed algorithm can correct the misalignment of prostate DWI and decrease the artifacts of ROI in the ADC maps. These ADC maps thus obtain sharper contours of lesions, which are helpful for improving the diagnosis and clinical staging of prostate cancer.

1. Introduction

Prostate cancer (PCa) is the second most frequently diagnosed cancer in men in western countries [1–4]. The early detection and treatment of PCa can decrease the death rate [5, 6]. The diagnosis of PCa is based on digital rectal examination, prostate specific antigen, transrectal ultrasonography, systematic transrectal biopsy, and diffusion-weighted magnetic resonance imaging (DWI) [7–10]. DWI depends on the microscopic mobility of water molecule diffusion [11–14]. Pathophysiological processes such as cancer are known to have an impact on cell density, which translates into different diffusion properties. Therefore, DWI can be applied in the detection, localization, and staging of PCa [15, 16]. The ADC map calculated from DWI is an important indicator in diagnosing PCa. The ADC proposed by Le Bihan is a noninvasive measure that provides quantitative information on the diffusion of water molecules in biological tissues [17–20], which explains why the ADC can be used to increase the sensitivity and specificity of the detection of PCa together with biopsy [21, 22].

The optimization of the ADC map brings great benefits in guiding targeted biopsy and in localizing and staging PCa, and it provides a roadmap for treatment planning and detecting residual or locally recurrent cancer after treatment [23, 24]. However, spatial misalignment is common in prostate DWI and leads to errors in the estimation of ADC.

Misalignments are particularly prone to occur in DWI because of motion and deformation, which can induce poor image quality and errors in the resulting ADCs. This situation is particularly serious in the case of elderly patients because of their weak respiratory control. Elderly patients cannot tolerate a supine position or remain still for a long time. In addition, the spatial alignment of the acquired DWI is not guaranteed if the pinched prostate deforms with peristalsis of the pelvic organs during the acquisition. The issue of image quality is commonly addressed by acquiring each DWI several times and averaging them. Despite

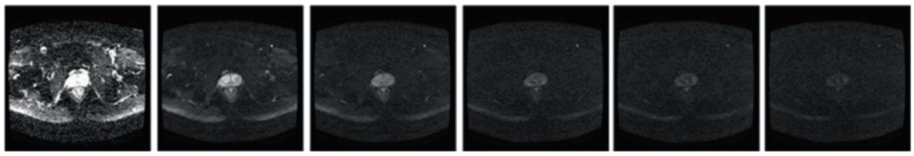


Figure 1: DWI acquired by $b = 0$, 500, 800, 1200, 1500, and 2000 s/mm^2 shows the tendency of intensity changes.

improving the signal-to-noise ratio (SNR) of the resulting ADC map, this technique does not compensate for spatial misalignment.

To the best of our knowledge, there are no correlative studies that address this misalignment in prostate DWI [15, 25–27]. To align the intraslice images and extract a more accurate ADC map, we present a semiautomatic registration algorithm, which enables us to correct the misalignments of DWI resulting from rigid body shifts, irregular distortions, and the patients' movements during the acquisition. The proposed algorithm evaluated 38 regions of interest (ROI) of 19 prostate DWI datasets. The ADC maps with our algorithm were then quantitatively compared with ADC maps without any image processing. The evaluation is based on the computation of similarity metric and reproducibility measures. The results showed that the visual quality of ADC maps can be improved through our proposed image registration algorithm.

2. Materials

2.1. Subjects. From August 2012 to January 2013, 19 male patients aged 68.96 ± 11.43 (50–87) years underwent conventional MRI and multi-b value DWI. It was found that there was a total of 20 PCa lesions and 15 benign prostatic hyperplasia (BPH) lesions. Their average prostate-specific antigen was 47.97 ± 72.98 (2.85–276.00) ng/ml. The mean Gleason score of tumors was 7.43 ± 1.09 (6–9). All patients underwent prostate 3T endorectal coil MRI followed by standard 14-core transrectal ultrasound-guided systematic (TRUS) biopsy. The experiments were conducted with the approval of the ethics committees of the participating institutions.

2.2. MR Images. The MR dataset obtained by the 3.0 Tesla (T) MR Scanner (Philips Achieva 3.0T; Philips Extended MR WorkSpace, Eindhoven, The Netherlands) contained T1-weighted imaging, T2-weighted imaging, DWI, and dynamic contrast-enhanced imaging [7]. DWI data were obtained using a multislice 2D echo-planar imaging (EPI) sequence in the transverse orientation. A single shot EPI sequence involved the following parameters: $b = 0$ s/mm^2 to 1200 s/mm^2, repetition time (TR) = 2000 ms, echo time (TE) = 68 ms, slice thickness = 3.0 mm, slice gap = 1 mm, flip angle = 90, field of view = 211 × 211 mm^2, and total effective scan time = 596 s. The size of the acquisition matrix was 160 × 156, and 16 slices were acquired with an in-plane spatial resolution of 1.31 × 1.31 mm^2, bandwidth 2985 Hz/pixel, and EPI factor 71. DWI sequences and ADC maps were reviewed by two radiologists independently and separately.

Patients with lesions suspicious for cancer on MRI underwent 14 or more core TRUS biopsies by the operator. All biopsies underwent centralized pathologic evaluation by a pathologist.

3. Methods

3.1. Algorithm Design: Global Transformation and Local Transformation (Steps 1, 2, and 3). Figure 1 shows an intraslice DWI sequence with a b value of up to 2000 s/mm^2. Image intensity dramatically reduces with an increase in the b value, which leads to prostate boundaries that are fuzzier and more difficult to segment. In addition, the interslice boundaries of the prostate are dynamic and irregular because of the different shooting positions. Thus, feature-based registration is not suitable for prostate DWI. Therefore, an intensity-based registration was applied in our scheme.

To correct the motion and deformation of DWI, intensity-based affine and nonrigid image registration algorithms have been developed [28–30]. The various steps are described in this section and illustrated in Figure 2.

To improve the accuracy of registration and reduce computing time, the center square region of the prostate is cropped from DWI (intraslice fixed position and uniform size) during image preprocessing (Figure 2, step 1). For the intraslice motion in the DWI sequence, Figure 3 shows the maximum interimage displacement of the prostate: 16 pixels (4.83%) horizontally and 29 pixels (8.9%) vertically. Therefore, we apply affine registration first to compensate for the intraslice motions between the images obtained in step 1, with $b \in \{50, 100, 150, 200, 500, 800\}$ s/mm^2 and $b = 0$ s/mm^2 (Figure 2, step 2).

The sequences obtained by affine registration are then placed into nonrigid registration (Figure 2, step 3). The original image will be retained if there are unreasonable rotations and scaling in the "affine" scheme. Thus, our algorithm consists of a global rigid transformation and a local nonrigid transformation:

$$T(x, y) = T_{\text{global}}(x, y) + T_{\text{local}}(x, y). \tag{1}$$

The image of $b = 0$ s/mm^2 in the sequence is chosen as a fixed reference image because it contains more details and has a higher SNR. The principle is to register other images ($b \neq 0$ s/mm^2) as float images to the fixed image (Figure 3).

The affine transformation model can describe the overall motion, scaling, and rotation of the prostate DWI obtained under the impact of different imaging times and magnetic field distortion. Therefore, in this study, we apply affine transformation to adjust these spatial misalignments as follows:

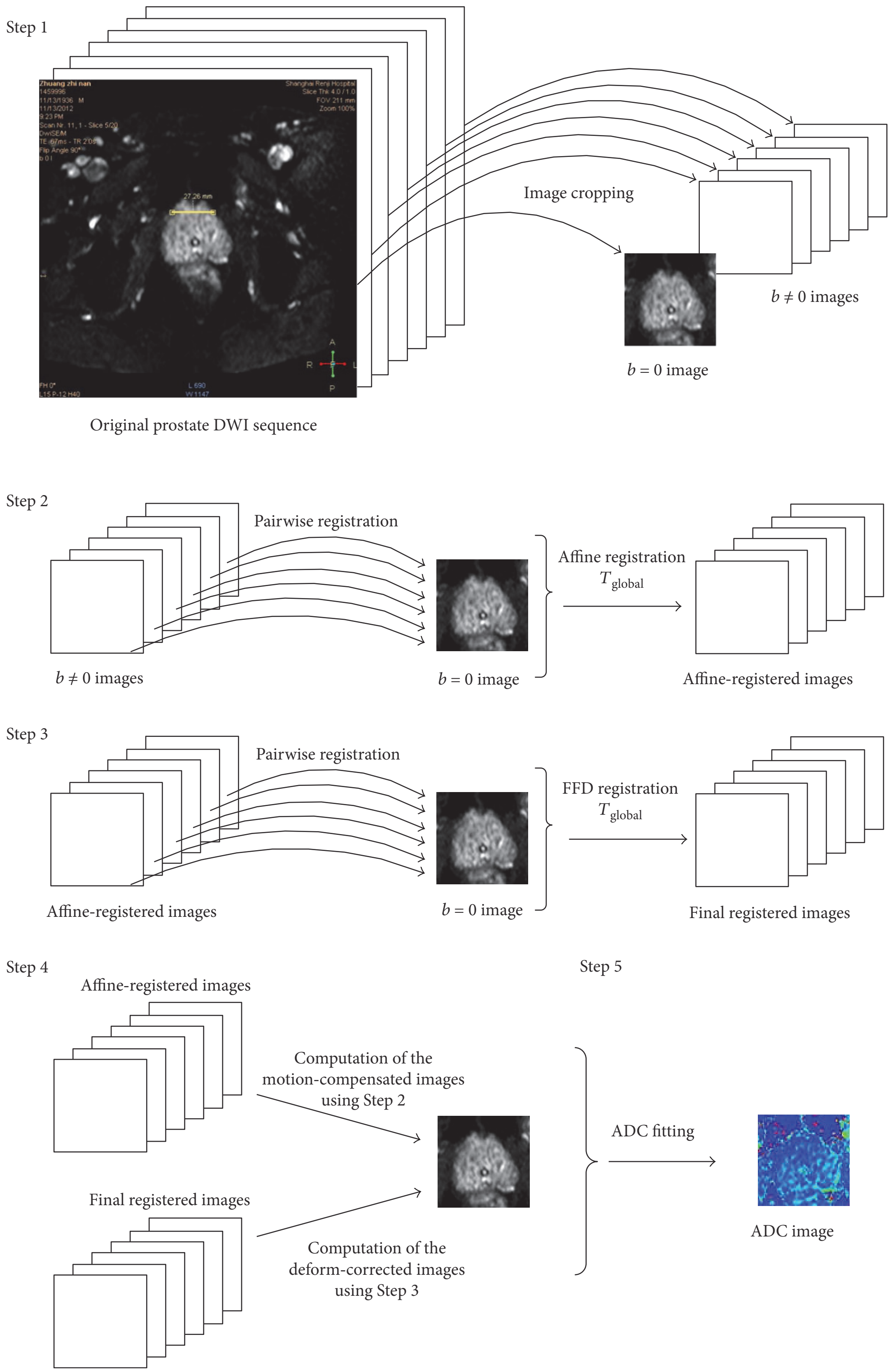


Figure 2: Main steps of the image registration. Step 1: image cropping. Steps 2 and 3: image registration. Step 4: visual and quantitative evaluation of registration. Step 5: ADC fitting.

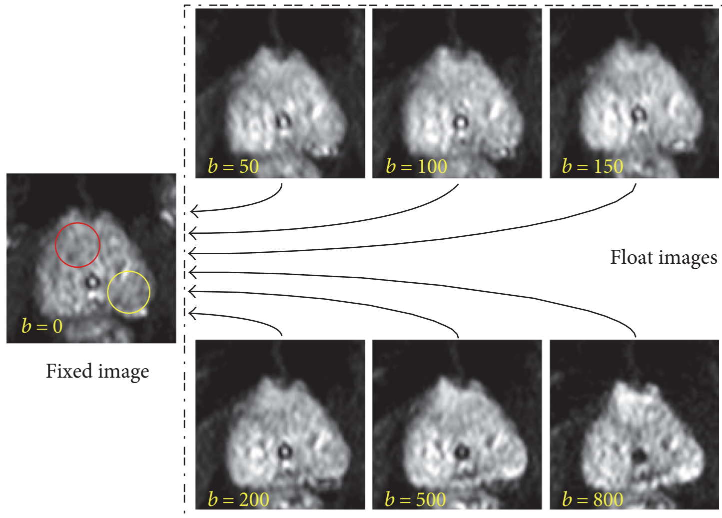


FIGURE 3: DWI sample acquired with $b = 0$, 50, 100, 150, 200, 500, and 800 in a 61-year-old patient. There was motion and deformation in the region of the prostate. For each patient, two types of elliptic ROIs are manually delineated on $b = 0\ \mathrm{s/mm^2}$. The ROI A (red circle) includes the nidus of the prostate, and the ROI B (yellow circle) covers the normal region. The ROIs are propagated differently in the "no processing," "affine," and "FFD" schemes.

$$T_{\text{global}}(x, y) = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & k_y \\ k_x & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}, \tag{2}$$

where Δx and Δy parameterize translation in the x and y directions, respectively; θ denotes rotation; s_x and s_y denote scaling; and k_x and k_y denote shearing in the x and y directions, respectively.

Since affine transformation can only roughly correct intraslice misalignment, local and fine transformation should be used to align the irregular intraimage regions of prostate DWI. The local deformation characteristics of the prostate are significantly different among different patients. It is difficult to describe these local deformations via parameterized transformations. Therefore, a free-form deformation (FFD) model based on cubic B-splines is necessary to correct the intra-aligned slices as a highly adaptive tool for modeling soft tissue deformation.

To define a spline-based FFD, we denote the domain of the image area as $\Omega = \{(x, y) | 0 \le x \le X, 0 \le y \le Y\}$. A $n_x \times n_y$ control mesh denoted by Φ is defined and applied to the image with uniform space δ. We denote the position of the control points on Φ as $\Phi_{i,j}$. Then, the FFD can be written as the 2-D tensor product of the familiar 1-D cubic B-splines:

$$T_{\text{local}}(x, y) = \sum_{l=0}^{3} \sum_{m=0}^{3} B_l(u) B_m(v) \Phi_{i+l,\, j+m}, \tag{3}$$

where $i = \lfloor x/\delta \rfloor - 1$, $j = \lfloor y/\delta \rfloor - 1$, $u = x/\delta - \lfloor x/\delta \rfloor$, $v = y/\delta - \lfloor y/\delta \rfloor$, and B_l and B_m represent the basic functions of lth and mth uniform cubic B-splines evaluated at u and v. They are defined as follows [28]:

$$\begin{aligned} B_0(u) &= \frac{(1-u)^3}{6}, \\ B_1(u) &= \frac{\left(3u^3 - 6u^2 + 4\right)}{6}, \\ B_2(u) &= \frac{\left(-3u^3 + 3u^2 + 3u + 1\right)}{6}, \\ B_3(u) &= \frac{u^3}{6}. \end{aligned} \tag{4}$$

FFD is implemented by manipulating an underlying mesh of control points to correct deformation. A deformation is defined on a sparse, regular grid of control points $\Phi_{i,j}$ placed over the float image and is then varied by defining the motion $g(\Phi_{i,j})$ of each control point. Using a spline interpolation kernel to compute the deformation values between the control points produces a locally controlled, globally smooth transformation.

The correction effect is affected by the resolution of the control mesh. The higher the resolution is, the finer the deformation correction is. However, since the cost of computation is intolerable, a hierarchal multiresolution that resembles the pyramid approach should be applied.

3.2. Optimization. An iterative optimization algorithm is performed with the pyramidal strategy, making this approach a coarse-to-fine strategy [26]. This approach has advantages such as higher convergence radius, more robust performance to local optimums, and faster speed. The motion artifacts are compensated for by a low-resolution strategy in nonrigid registration. Then, the resolution is increased so the local deformation is aligned. The quasi-newton minimization package is applied in local models to reduce the time required to compute the cost function (see (3)) until the termination criteria are satisfied or a maximum number of 800 iterations per resolution is reached [30]. For

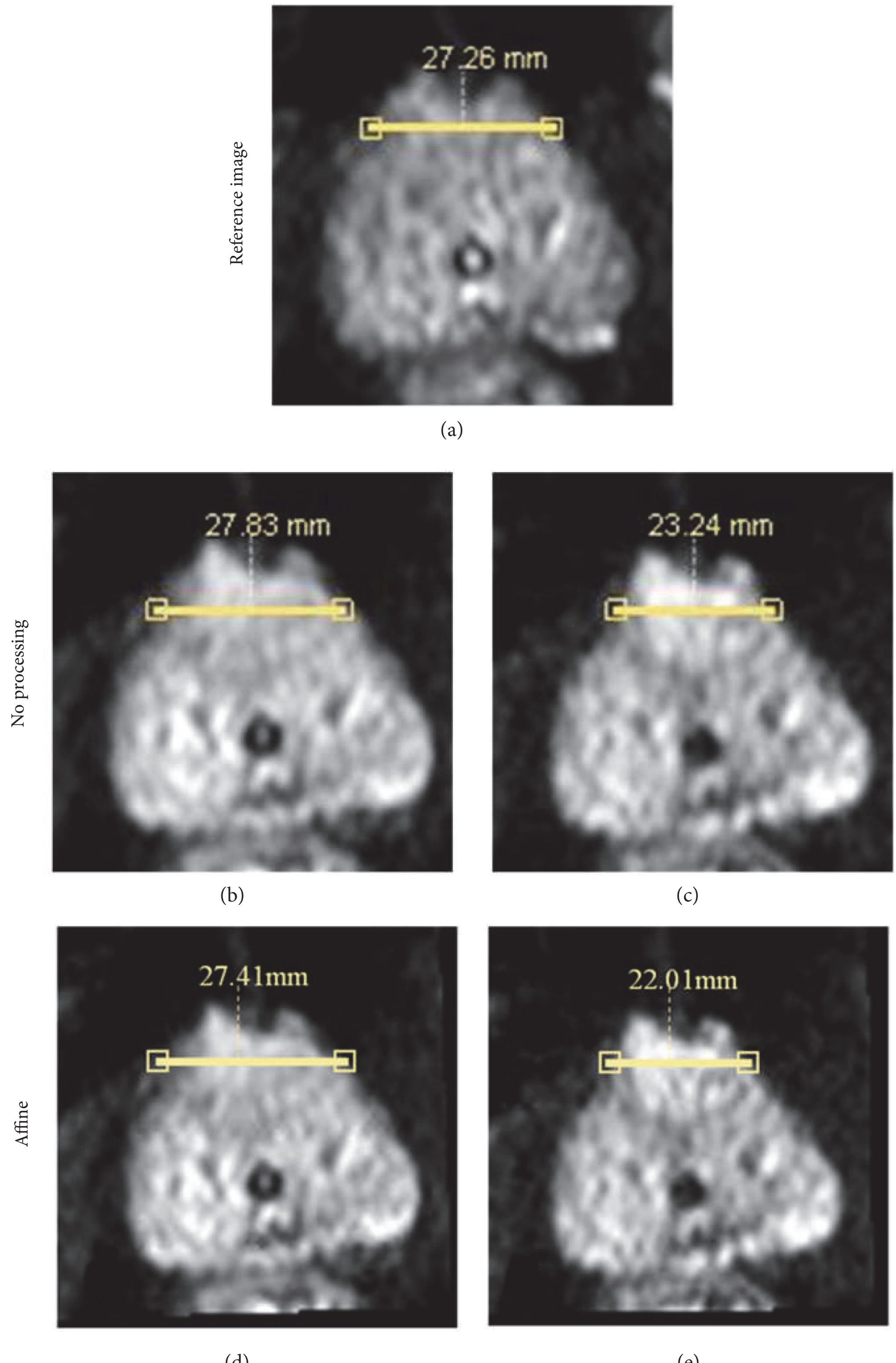
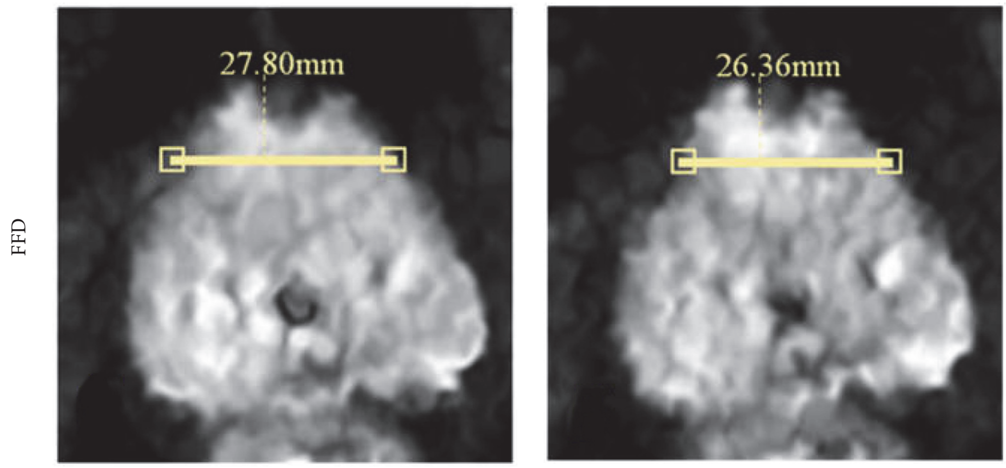


Figure 4: Continued.



(f) (g)

Figure 4: (a) $b = 0\ \mathrm{s/mm^2}$ DW image is the reference image with no processing (case 19). (b) and (c) are the respective samples of $b = 500\ \mathrm{s/mm^2}$ and $b = 800\ \mathrm{s/mm^2}$ DW images in the "no processing" schemes. (d) and (e) are the $b = 500\ \mathrm{s/mm^2}$ and $b = 800\ \mathrm{s/mm^2}$ DW images in the "affine" schemes. (f) and (g) with FFD registration are mostly visually similar to (a).

optimization, the ant colony algorithm minimization method is applied to affine transformation.

3.3. Similarity Measure (Step 4). The most commonly used similarity measures are based on intensity differences, intensity cross-correlation, and information theory, such as normalized cross-correlation (NCC), NMI, and mean-squared error (MSE) [26]. The intensity distribution relationship of DWI is linear. Therefore, NCC is adopted as our cost function while performing the actual registration in this study. The NCC of the fixed image and the float image is defined by

$$\mathrm{NCC}(I_1, I_2) = \frac{1}{N-1}\sum_x \sum_y \frac{\left(I_1(x,y) - \overline{I_1}\right)\left(I_2(x,y) - \overline{I_2}\right)}{\sigma_{I_1}\sigma_{I_2}}, \quad (5)$$

where I_1 is the fixed image, I_2 is the float image, and σ_{I_1} and σ_{I_2} are the standard deviations of I_1 and I_2, respectively. NMI is chosen as a similarity metric to evaluate motion and deformation compensation.

3.4. ADC Computation (Step 5). The ADC values are then extracted by curve fitting from the intraimage registration DWI (see (6)). ADCs are computed from DWI characterized by different diffusion weighting factors (b values). Several models to evaluate ADC have been developed: monoexponential model, a biexponential model with two independent fractions, a statistical model with a distribution fraction, and the recent kurtosis model [22]. In this paper, we calculated the ADC with the monoexponential model as follows:

$$S(b) = S_0 e^{-\mathrm{ADC}_{\mathrm{mono}} b}, \quad (6)$$

where $S(b)$ is the signal intensity and depends on the strength of diffusion weighting characterized by the b factor (b). $\mathrm{ADC}_{\mathrm{mono}}$ is the diffusion coefficient, and S_0 is the signal without applying a diffusion weighting gradient ($b = 0\ \mathrm{s/mm^2}$).

4. Experiments and Results

4.1. Background and Considered Schemes. Nineteen patient datasets were obtained from Renji Hospital. A prostate DWI include 16 sequences, according to different b values, and each sequence contained 6–8 images. Each sample was saved in a DICOM format image that contained information on the b value and flip angle. A total of 260 transformations were performed, including affine and nonrigid registration.

The effect of registration is evaluated by comparing DWI visualization, NMI, and ADC maps in three schemes (no processing, affine, and FFD). The first scheme is called "no processing," which means that the ADC curve is obtained with the original sequence. The second scheme is referred to as "affine." It involves applying translation and zoom to each of the acquired DWIs before obtaining the ADC curve. The third scheme, denoted "FFD," is that the free-form deformation registration is implemented in the sequence before the ADC curve is fitted.

4.2. Visual Evaluation of Image Registration. The first experiment seeks to verify the validity of the registration and evaluate the registration effect visually. The specific approach is to use the Philips DICOM Viewer to display the DWI, and the radiologist manually selects the pronounced deformation cases by observing all of the samples. After this step, the radiologist uses the "line function" of this software to measure the width or height of the prostate and then estimate the degree of deformation. The last step is to compare the visual effects of the first scheme with the third one and evaluate the degree of deformation compensation. The experimental results are shown in Figure 4.

As shown in Figure 4, the alignments of $b = 500\ \mathrm{s/mm^2}$ DW image, $b = 800\ \mathrm{s/mm^2}$ DW image, and $b = 0\ \mathrm{s/mm^2}$ DW image can be compared in the "no processing" and "FFD" schemes (case 19). Figure 4 shows that DW image of $b = 500\ \mathrm{s/mm^2}$ is similar to that of $b = 0\ \mathrm{s/mm^2}$. Obviously, there is visual extrusion deformation in the prostate image

Table 1: NMI values of ROI A.

	b values				
	50	100	200	500	800
No processing	1.3883 ± 0.3239	1.3649 ± 0.4168	1.1993 ± 0.3255	1.0995 ± 0.3354	0.9230 ± 0.3511
Affine	1.4051 ± 0.3136	1.3285 ± 0.2773	1.2353 ± 0.3226	1.1405 ± 0.3375	0.9945 ± 0.3179
FFD	2.2686 ± 0.3576	2.1717 ± 0.3192	2.0518 ± 0.3595	1.9412 ± 0.3756	1.8141 ± 0.4151

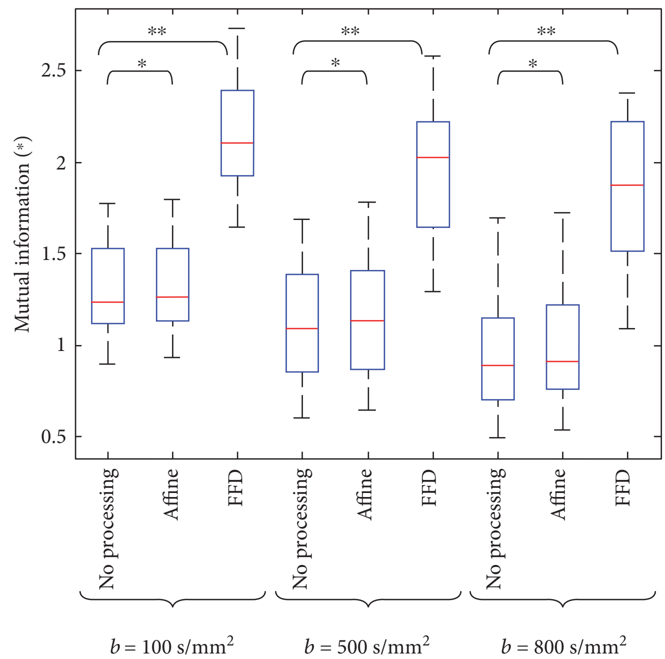


Figure 5: NMI with $b = 100$, 500, and 800 s/mm^2 images and the $b = 0$ s/mm^2 images. The highest NMI is obtained for the "FFD" scheme. Paired t-tests between the schemes: $^{*}P < 0.05$, $^{**}P < 0.01$.

of $b = 800$ s/mm^2. However, this kind of image deformation can be compensated effectively in the FFD scheme. We also observe that few unrealistic artifacts are generated by FFD registration with different b value images, indicating that the adopted NCC is adequate.

4.3. Quantitative Evaluation of Motion and Deformation Compensation. The second experiment is dedicated to the quantization of alignment accuracy in the three schemes. Therefore, NMI of ROI is computed in compensation accuracy evaluation experiments. There are two ROIs per PCa patient, the lesion area located in the red circle and the general area located in the yellow circle, as shown in Figure 3. For each sample, two ROIs are defined with a diameter of 16 mm on the first $b = 0$ s/mm^2 image in the "no processing" scheme. The same ROIs are used in the "affine" and "FFD" schemes. The ROI circle encompasses approximately 120 pixels. The ROI A (Figure 2) is manually positioned in the heterogeneous tumor region of the prostate. To compare the effect of the homogeneous region, the ROI B is selected in a common area as a reference for ROI A.

Table 1 reports the NMI values of each b value image to $b = 0$ s/mm^2 image in all samples of the three schemes, which increased by 2.6% on average after affine transformation and 73.32% after nonrigid registration.

Furthermore, to compare the compensation accuracy of the low b value ($b = 100$ s/mm^2) image subset with the high b value ($b = 500, 800$ s/mm^2) image subsets, in the three schemes, the paired t-test is implemented. The subsets' statistics results are shown in Figure 5.

In terms of NMI, Figure 5 shows that the aligned correlation in the "FFD" scheme is higher than that in the "no processing" and "affine" schemes with the respective mean NMI values of 1.3146, 1.3346, and 2.1696 for $b = 100$ s/mm^2, 1.1284, 1.1627, and 1.9675 for $b = 100$ s/mm^2, and 0.9482, 1.0133, and 1.8420 for $b = 800$ s/mm^2. NMI tends to decrease gradually with increasing b values in the three schemes. The results show that NMI values are significantly different between the "no processing" and "FFD" schemes.

4.4. Quantitative Analysis of ADC Values. The third experiment seeks to compare the ADC values across the three schemes and evaluate the effect of the registration. The ADC median, homogeneity (using interquartile ranges (IQRs) [31]), mean level of diffusion (mean value) of ROIs, and the reproducibility (using the standard deviation (STD)

Table 2: Mean value, median value, and interquartile ranges (IQRs) of ADC values of ROI A.

		No processing	Affine	Nonrigid			No processing	Affine	Nonrigid
	ADC median	1.0223	1.0210	1.0254		IQR	1.6132	1.6954	1.5387
Case 1	Mean	1.0783	1.0165	1.0312	Case 11	ADC median	0.6351	0.6318	0.5758
	IQR	1.0122	1.0166	1.0291		Mean	1.0238	1.4928	1.1099
	ADC median	0.8718	0.8643	0.8333		IQR	1.0282	1.5977	1.1027
Case 2	Mean	0.95283	0.95735	1.0243	Case 12	ADC median	0.6161	0.5463	0.4804
	IQR	0.94842	0.95098	1.0026		Mean	0.9565	1.1463	0.7726
	ADC median	0.7015	0.7474	0.7198		IQR	0.9780	1.1532	0.8041
Case 3	Mean	0.9460	1.0262	1.0247	Case 13	ADC median	0.9387	0.9678	0.8982
	IQR	1.0927	1.3608	0.9896		Mean	1.2657	1.3359	1.3019
	ADC median	1.2350	0.9852	0.9813		IQR	1.1897	1.2429	1.2441
Case 4	Mean	1.1309	0.9509	0.9042	Case 14	ADC median	1.0832	1.1090	0.4436
	IQR	1.3482	1.2789	1.2425		Mean	1.7366	1.7929	0.9207
	ADC median	1.0299	1.0054	0.9795		IQR	2.0324	2.1214	0.9971
Case 5	Mean	1.1452	1.0806	1.0023	Case 15	ADC median	1.2260	1.1649	1.0726
	IQR	0.8864	0.8529	0.7936		Mean	1.9196	1.9111	1.6321
	ADC median	1.0833	0.9710	0.9115		IQR	2.5168	2.4777	2.1027
Case 6	Mean	1.0994	1.0265	1.1976	Case 16	ADC median	1.0859	1.0600	1.0531
	IQR	0.8550	0.89675	0.81655		Mean	1.2464	1.3655	1.2247
	ADC median	0.8444	0.8440	0.8330		IQR	0.9431	0.9959	0.7185
Case 7	Mean	1.1060	1.1296	1.1780	Case 17	ADC median	0.8975	0.8090	0.6590
	IQR	1.5310	1.5939	1.5410		Mean	1.3103	1.6779	1.0683
	ADC median	1.7388	1.7637	1.7112		IQR	1.2911	1.9409	0.9682
Case 8	Mean	1.6751	1.6630	1.7845	Case 18	ADC median	0.9277	0.9185	0.8545
	IQR	1.5764	1.5617	1.5488		Mean	1.4199	0.9185	1.2287
	ADC median	0.7816	0.7363	0.6901		IQR	1.2304	1.2366	1.0727
Case 9	Mean	0.7934	0.7469	0.6692	Case 19	ADC median	0.3949	0.4006	0.2846
	IQR	0.6996	0.6860	0.6778		Mean	0.9149	1.0515	0.8216
Case 10	ADC median	0.9801	0.9200	0.9302		IQR	0.9547	1.1746	0.7876
	Mean	1.2127	1.2478	1.2043					

All values are given in $\mu m^2/ms$.

of ADCs) across the three schemes are analyzed quantitatively with all samples. Paired t-tests are used to compare the distributions of the median ADC values, IQR, and mean values obtained in the three schemes.

The mean value, median value, and IQR computed from the ROI A of all datasets are reported in Table 2 and shown graphically in Figure 6. For a given dataset, the above quantitative parameters calculated from the two ROIs are quite similar. For all datasets, these parameters have a normal distribution. The main results corresponding to the average of the median ADCs, mean values, and IQRs of two ROIs are presented in Table 3.

The median ADCs of ROI A are 0.9523 in the "no processing" scheme, 0.9193 in the "affine" scheme, and 0.8385 in the "FFD" scheme. For ROI B, the respective median ADCs are 1.0707, 1.0335, and 0.9427. Statistics suggest that for ROI A, the median ADCs obtained in the "no processing" scheme are always higher compared with the "affine" (3.6%) and "FFD" (13.6%) schemes. The same trend occurs in ROI B. The respective mean ADCs of ROI A are 1.2070, 1.2388, and 1.1000 in the three schemes. For ROI B, the mean ADCs are 1.3543, 1.3900, and 1.2342, respectively. For both ROIs, the lowest mean ADCs are obtained in the "FFD" scheme, which indicates that the number of error values is partly reduced. Given a dataset, IQRs (characterizing the homogeneity of the ADCs) within both ROIs are lower with FFD registration than when "no processing" or "affine" transformation is applied to the images. Table 4 shows that STDs are in general reduced in the "FFD" scheme, with respect to the "no processing" and "affine" schemes.

4.5. Visual Evaluation of ADC Map. ADC maps are evaluated visually by radiologists in the three schemes. An example of ADC maps fitted with the 19th database is shown in Figure 7. Visual inspection of the ADC maps indicates that the "FFD" scheme improves the visual quality of the ADC maps with respect to the "no processing" scheme and the "affine" scheme: the suspected nidus areas are better visualized and the number of pixels for which the fitting fails to decrease. The ADC maps fitted with no processing sequence and affine transformation sequence visually appear to be similar.

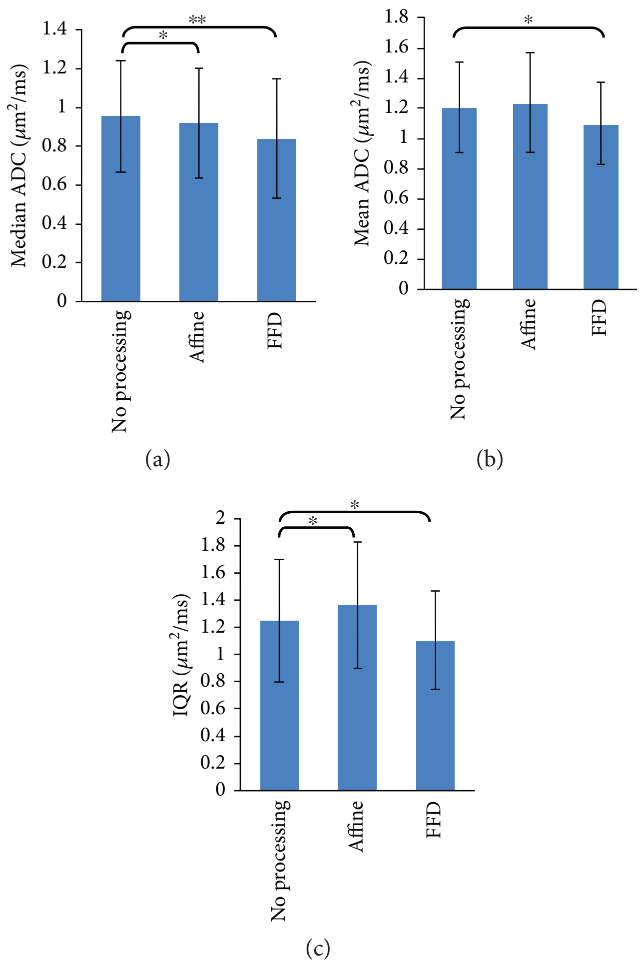


FIGURE 6: (a) Median ADCs, (b) mean ADCs, and (c) interquartile range (IQR) of ROI A are averaged over the 19 samples. The error bars represent the standard deviation of the measurements over 19 samples. Paired t-tests are used to compare the median ADC, mean ADC, and IQR datasets obtained in the "affine" and "FFD" schemes respect with the "no processing" scheme. $^*P < 0.05$, $^{**}P < 0.01$. Median ADCs, mean ADCs, and IQR obtained the lowest value in the "FFD" scheme.

TABLE 3: Average of the median ADCs, mean values, and IQRs.

		No processing	Affine	Nonrigid
(a)				
Mean ± STD	ADC median	0.9523 ± 0.2863	0.9193 ± 0.2820	0.8385 ± 0.3057
	Mean	1.2070 ± 0.2983	1.2388 ± 0.3290	1.1000 ± 0.2708
	IQR	1.2488 ± 0.4483	1.3597 ± 0.4647	1.1041 ± 0.3631
(b)				
Mean ± STD	ADC median	1.0707 ± 0.1632	1.0335 ± 0.3581	0.9427 ± 0.3527
	Mean	1.3543 ± 0.1650	1.3900 ± 0.3165	1.2342 ± 0.1742
	IQR	1.4812 ± 0.4368	1.5256 ± 0.3238	1.3388 ± 0.3554

All values are given in μm^2/ms. The average lines contain mean and standard deviation values calculated with all of 19 datasets. (a) ROI A, (b) ROI B.

The ADC map obtained by 7 b values ranging from 0 to 800 s/mm^2 is shown in Figure 7(a). The affine and nonrigid registration results are shown in Figures 7(b) and 7(c), respectively. Alignment was assessed by two experienced radiologists who specialize in both T2WI and DWI. The local suspicious increase of intensity is visible in the high b value DW images. This region presents a diffusion-limited change, which was detected in the corresponding ADC map shown in Figure 7(a). However, the border of lesions is less clear in this ADC map fitted with the original images. Figure 7(b) fitted with affine registration images displays the location and boundaries of the lesion more clearly. It also shows that the artifacts produced by deformation cannot be eliminated completely. In contrast, Figure 7(c) fitted with nonrigid registration images shows obvious diffusion-limited changes that are highly suspected of being PCa. It provides more accurate reference information for aspiration biopsy. This is also confirmed by the gold standard [8] for the diagnosis of prostate cancer. The lesion is confirmed to be PCa by TRUS biopsy with a Gleason score of $4 + 4 = 8$.

Table 4: STD of ADCs of ROI A.

	Case1	Case2	Case3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case10
No processing	0.6649	0.6205	0.7803	0.7346	0.6961	0.5821	0.8673	0.9296	0.4608	0.8952
Affine	0.6725	0.6179	0.8598	0.6989	0.7861	0.5980	0.8955	0.9325	0.4781	0.9260
FFD	0.5159	0.4314	0.4464	0.5537	0.5358	0.4364	0.6457	0.8458	0.3066	0.8344
	Case11	Case12	Case13	Case14	Case15	Case16	Case17	Case18	Case19	
No processing	1.1471	1.1374	1.1949	1.8482	1.8605	0.9684	1.3349	1.5329	1.3889	
Affine	1.5842	1.3482	1.2936	1.8832	1.8377	1.0537	1.5392	1.5870	1.5918	
FFD	1.0058	1.1089	1.0955	1.3184	1.6354	0.9447	1.2388	1.3291	1.1266	

All values are given in $\mu m^2/ms$.

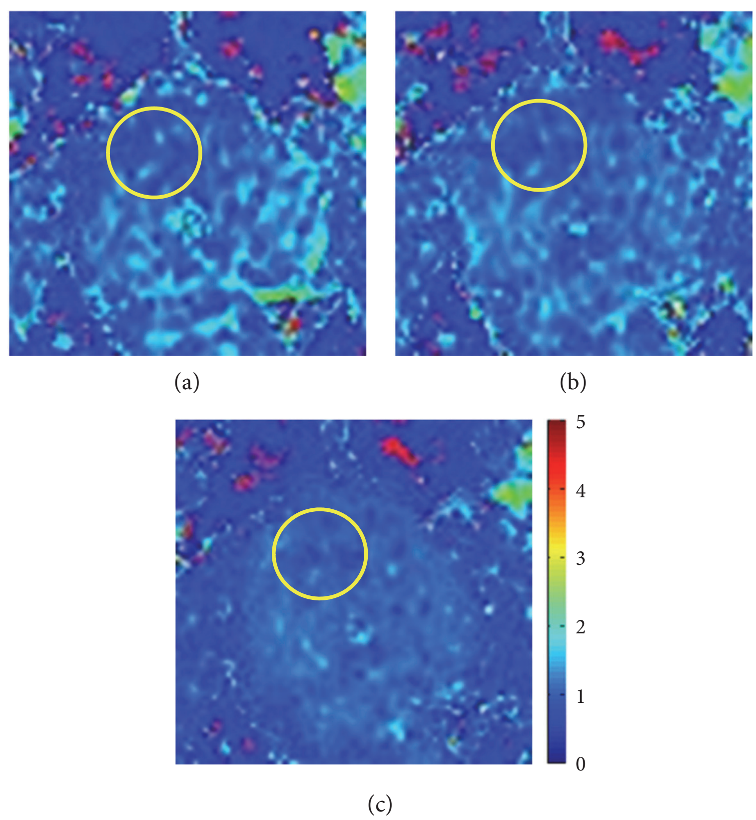


Figure 7: ADC maps of the 19th sample fitted with DWI. Color scales are given in $\mu m^2/ms$. (a) ADC map fitted with original DWI; (b) ADC map fitted with affine transformation DWI; (c) ADC map fitted with nonrigid registration DWI.

The above evaluation and conclusion of the two radiologists illustrate that image registration is effective for addressing the deformation and motion of prostate DWI.

5. Discussion and Conclusion

In this study, we proposed a semiautomatic algorithm for the registration of prostate DWI to improve the ADC map. The influence of this image registration algorithm on ADC quantification was investigated in three schemes: no processing, with affine transformation, and with FFD registration of acquired DWI.

For the method of addressing the misalignment in DWI sequences, we apply affine transformation to eliminate interimage sliding motions and then use the nonrigid registration to correct intraimage deformations. This method with a global affine transformation is more time-efficient. NMI as the similarity measure is computed to elevate the registration result in both the "affine" and "FFD" schemes. The increase of NMI in the nonrigid registration scheme is distinct from that in the "affine" scheme, which indicates that the DWI is better aligned in the "FFD" scheme. It also shows that affine transformation is not sufficient for the alignment of prostate DWI and that the nonrigid registration with deformation compensation is compulsory. In addition, we assume that the dynamic SNR (Figure 2) with different b values in DWI sequence would challenge the validity of our algorithm. Therefore, low $(b = 100\ s/mm^2)$

and high b value ($b = 500, 800\ \mathrm{s/mm^2}$) subsets extracted from the acquired DWI are investigated. However, statistical results (Figure 4) show that our algorithm still performs well for different subsets. That is, our registration algorithm has good robustness.

The "no processing" and "affine" schemes result in the overestimation of the mean ADC values for ROI A of all datasets, by 9.7% and 12.6%, respectively, and the IQRs are overestimated by 13.1% and 23.1%, respectively, compared with FFD scheme. The median ADC in the "no processing" scheme is overestimated by 3.59% and 13.6%, respectively, compared with that in the "affine" scheme and "FFD" scheme. Variability measure (STD) is generally improved with FFD registration. Similar trends also appear in ROI B for median and mean ADCs. However, IQR values of ROI B have no significant difference with or without registration. This indicates that the registration algorithm can make the lesion boundaries clearer in the inhomogeneous region (ROI A). These effects are visually confirmed in Figure 7. These results indicate that FFD registration significantly improves the visual quality of ADC maps. It also suggests that the effect of registration on the ADC maps is mainly determined by the deformation compensation of the prostate. This compensation of deformation is influenced definitively by the FFD grid resolution: a higher resolution would obtain a better effect. However, higher resolution is time-consuming, so the highest resolution in this paper is limited to within 90×90. The experimental results indicate that the algorithm with this resolution is capable to be sufficiently accurate.

The radiologists observed and analyzed the malignant regions while reviewing these T2-weighted images and DWI and ADC maps. They found that the ADC maps are valid in the "affine" and "FFD" schemes. The ADC maps of the "FFD" scheme provide the most details for the clinical staging of PCa. Therefore, the ADC maps of the "FFD" scheme are accepted by radiologists and correspond with the clinical diagnosis. The registration algorithm can eliminate the drift and deformation of the acquired DWI. This indicates that our method in this study is efficient.

In conclusion, this study shows that image registration can correct the misalignment of prostate DWI and improve ADC maps. The clinicians believe that the ADC maps obtained from the registered DWI are more effective for the detection, diagnosis, and staging of PCa. For future work, it could be interesting to compare this method with the use of a nonrigid 3-D transformation model, correcting for both intraslice and interslice motions.

Conflicts of Interest

The authors declare that they have no conflict of interest to report with respect to this paper.

Acknowledgments

The authors thank Dr. Lixu Gu for his assistance in acquiring the images. The authors would also like to thank Drs. Jianrong Xu and Lianming Wu, consultant radiologists at Renji Hospital, for their enthusiastic support. This work is supported by the National Natural Science Foundation of China (no. 11271106), the Scientific Programs for Hebei Higher Education Institutions of China (no. Z2015147), and the Youth Foundation from the Educational Department of Hebei Province (no. QN2016169).

References

[1] C. Cuello, U. Saragovi, P. D. Ruisseau, P. Gold, N. Bernard, and S. Moffett, "Prostate cancer diagnosis and treatment," U.S. Patent US8512702 B2, 2013.

[2] S. Rebecca, M. Jiemin, Z. Zhaohui, and J. Ahmedin, "Cancer statistics, 2014," *CA Cancer Journal for Clinicians*, vol. 64, no. 1, pp. 9–29, 2014.

[3] R. Siegel, D. Naishadham, and A. Jemal, "Cancer statistics, 2013," *CA Cancer Journal for Clinicians*, vol. 63, no. 1, pp. 11–30, 2013.

[4] M. Malvezzi, P. Bertuccio, F. Levi, V. C. La, and E. Negri, "European cancer mortality predictions for the year 2013," *Annals of Oncology Official Journal of the European Society for Medical Oncology*, vol. 24, no. 3, pp. 947–956, 2013.

[5] C. Eheman, A. G. Zauber, R. N. Anderson et al., "Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates," *Cancer*, vol. 116, no. 3, pp. 544–573, 2010.

[6] R. Etzioni, A. Tsodikov, A. Mariotto et al., "Quantifying the role of PSA screening in the US prostate cancer mortality decline," *Cancer Causes & Control*, vol. 19, no. 2, pp. 175–181, 2008.

[7] N. Mottet, P. J. Bastian, and J. Bellmunt, "European Association of Urology guidelines on prostate cancer 2015," *European Association of Urology*, pp. 22–24, 2015.

[8] A. Heidenreich, J. Bellmunt, M. Bolla et al., "Guidelines on prostate cancer," *Drukkerij Gelderland Bv*, vol. 40, no. 2, pp. 546–551, 2013.

[9] T. Barrett, B. Turkbey, and P. L. Choyke, "PI-RADS version 2: what you need to know," *Clinical Radiology*, vol. 70, no. 11, pp. 1165–1176, 2015.

[10] M. Quentin, D. Blondin, J. Klasen et al., "Comparison of different mathematical models of diffusion-weighted prostate MR imaging," *Magnetic Resonance Imaging*, vol. 30, no. 10, pp. 1468–1474, 2012.

[11] H. Dijkstra, P. Baron, P. Kappert, and M. Oudkerk, "Effects of microperfusion in hepatic diffusion weighted imaging," *European Radiology*, vol. 22, no. 4, pp. 891–899, 2012.

[12] A. Andreou, D. M. Koh, D. J. Collins et al., "Measurement reproducibility of perfusion fraction and pseudodiffusion coefficient derived by intravoxel incoherent motion diffusion-weighted MR imaging in normal liver and metastases," *European Radiology*, vol. 23, no. 2, pp. 428–434, 2013.

[13] D. Lebihan, "Intravoxel incoherent motion perfusion MR imaging: a wake-up call," *Radiology*, vol. 249, no. 3, pp. 748–752, 2008.

[14] N. E. Larsen, S. Haack, L. P. S. Larsen, and E. M. Pedersen, "Quantitative liver ADC measurements using diffusion-weighted MRI at 3 Tesla: evaluation of reproducibility and perfusion dependence using different techniques for respiratory compensation," *Magma Magnetic Resonance Materials in Physics Biology & Medicine*, vol. 26, no. 5, pp. 431–442, 2013.

[15] S. Wang, K. Burtt, B. Turkbey, P. Choyke, and R. M. Summers, "Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research," *BioMed Research International*, vol. 2014, Article ID 789561, 11 pages, 2014.

[16] K. Dow-Mu, D. J. Collins, and M. R. Orton, "Intravoxel incoherent motion in body diffusion-weighted MRI: reality and challenges," *American Journal of Roentgenology*, vol. 196, no. 6, pp. 1351–1361, 2011.

[17] J. L. Leake, R. Hardman, V. Ojili et al., "Prostate MRI: access to and current practice of prostate MRI in the United States," *Journal of the American College of Radiology*, vol. 11, no. 2, pp. 156–160, 2014.

[18] Y. Miyamoto and K. Jinnai, "MR imaging of the prostate in clinical practice," *Brain Research*, vol. 21, no. 6, pp. 379–392, 2008.

[19] T. Hambrock, D. M. Somford, H. J. Huisman et al., "Relationship between apparent diffusion coefficients at 3.0-T MR imaging and Gleason grade in peripheral zone prostate cancer," *Radiology*, vol. 259, no. 2, pp. 453–461, 2011.

[20] D. Le Bihan, E. Breton, D. Lallemand, P. Grenier, E. Cabanis, and M. Laval-Jeantet, "MR imaging of incoherent motion: application to diffusion and perfusion in neurologic disorders," *Radiology*, vol. 161, no. 2, pp. 401–407, 1986.

[21] E. M. Charlesedwards and N. M. Desouza, "Diffusion-weighted magnetic resonance imaging and its application to cancer," *Cancer Imaging the Official Publication of the International Cancer Imaging Society*, vol. 6, no. 1, pp. 135–143, 2006.

[22] J. F. A. Jansen, H. E. Stambuk, J. A. Koutcher, and A. Shukla-Dave, "Non-Gaussian analysis of diffusion-weighted MR imaging in head and neck squamous cell carcinoma: a feasibility study," *American Journal of Neuroradiology*, vol. 31, no. 4, pp. 741–748, 2010.

[23] M. De Luca, V. Giannini, A. Vignati et al., "A fully automatic method to register the prostate gland on T2-weighted and EPI-DWI images," *International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC*, pp. 8029–8032, Boston, MA, 2011.

[24] M. A. Bjurlin, X. Meng, J. Le Nobin et al., "Optimization of prostate biopsy: the role of magnetic resonance imaging targeted biopsy in detection, localization and risk assessment," *Journal of Urology*, vol. 192, no. 3, pp. 648–658, 2014.

[25] D. Francisco, D. N. Costa, Y. Qing, N. M. Rofsky, R. E. Lenkinski, and P. Ivan, "Geometric distortion in diffusion-weighted MR imaging of the prostate-contributing factors and strategies for improvement," *Academic Radiology*, vol. 21, no. 6, pp. 817–823, 2014.

[26] F. P. M. Oliveira and J. O. M. R. S. Tavares, "Medical image registration: a review," *Computer Methods in Biomechanics & Biomedical Engineering*, vol. 17, no. 2, pp. 73–93, 2014.

[27] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: a survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[28] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[29] A. Valsecchi, S. Damas, and J. Santamaria, "Evolutionary intensity-based medical image registration: a review," *Current Medical Imaging Reviews*, vol. 9, no. 4, pp. 283–297(215), 2013.

[30] J. P. Pluim, J. B. Maintz, and M. A. Viergever, "F-information measures in medical image registration," *Surgery*, vol. 136, no. 12, pp. 1508–1516, 2004.

[31] J. M. Guyader, M. D. Livia Bernardin, N. H. M. Douglas, D. H. J. Poot, W. J. Niessen, and S. Klein, "Influence of image registration on apparent diffusion coefficient images computed from free-breathing diffusion MR images of the abdomen," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 2, pp. 315–330, 2014.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 5703216, 7 pages
https://doi.org/10.1155/2017/5703216

Research Article

Segmentation Method for Magnetic Resonance-Guided High-Intensity Focused Ultrasound Therapy Planning

A. Vargas-Olivares,[1] O. Navarro-Hinojosa,[2] M. Maqueo-Vicencio,[2] L. Curiel,[3] M. Alencastre-Miranda,[2] and J. E. Chong-Quero[1]

[1]*Tecnologico de Monterrey, Campus Estado de México, Atizapán de Zaragoza, MEX, Mexico*
[2]*Tecnologico de Monterrey, Campus Santa Fe, Álvaro Obregón, Ciudad de México, Mexico*
[3]*Electrical Engineering Department, Lakehead University, Thunder Bay, ON, Canada*

Correspondence should be addressed to A. Vargas-Olivares; a00457284@itesm.mx

Received 24 February 2017; Accepted 26 April 2017; Published 22 June 2017

Academic Editor: Junfeng Gao

High-intensity focused ultrasound (HIFU) is a minimally invasive therapy modality in which ultrasound beams are concentrated at a focal region, producing a rise of temperature and selective ablation within the focal volume and leaving surrounding tissues intact. HIFU has been proposed for the safe ablation of both malignant and benign tissues and as an agent for drug delivery. Magnetic resonance imaging (MRI) has been proposed as guidance and monitoring method for the therapy. The identification of regions of interest is a crucial procedure in HIFU therapy planning. This procedure is performed in the MR images. The purpose of the present research work is to implement a time-efficient and functional segmentation scheme, based on the watershed segmentation algorithm, for the MR images used for the HIFU therapy planning. The achievement of a segmentation process with functional results is feasible, but preliminary image processing steps are required in order to define the markers for the segmentation algorithm. Moreover, the segmentation scheme is applied in parallel to an MR image data set through the use of a thread pool, achieving a near real-time execution and making a contribution to solve the time-consuming problem of the HIFU therapy planning.

1. Introduction

High-intensity focused ultrasound (HIFU) is a minimally invasive therapy modality in which ultrasound beams are concentrated at a focal region, producing a rise of temperature and selective ablation within the focal volume and leaving surrounding tissues intact [1]. HIFU has been proposed for the safe ablation of both malignant and benign tissues and as an agent for drug delivery [2]. Magnetic resonance imaging (MRI) has been proposed for guidance and monitoring for the therapy as it provides anatomical images with an adequate spatial resolution. On the other hand, MRI is sensitive to temperature changes [3]. The combination of HIFU and MRI is known as magnetic resonance-guided HIFU (MRgHIFU). The objects that are found in a HIFU treatment include the ultrasonic transducer, acoustic coupling medium (such as water, oil, and gel pads), and the tissue to be treated as shown in Figure 1.

In MRgHIFU therapy planning, the identification of regions of interest, such as regions within the tissue and around the transducer, is performed in order to define the target tissue region and the transducer position along with the localization of its geometric focus. Several MR images are used to cover the volume of interest to be treated. Once the position of the geometric focus (focal point) is obtained, it is guided towards the target tissue [4]. If a proper identification is achieved, it would be possible to calculate in advance the effects of the application of the therapy, given the current distribution of regions of interest. Image segmentation algorithms have been proposed as an alternative to the manual identification of the regions of interest, a time-consuming problem in the therapy planning [5]. This problem becomes

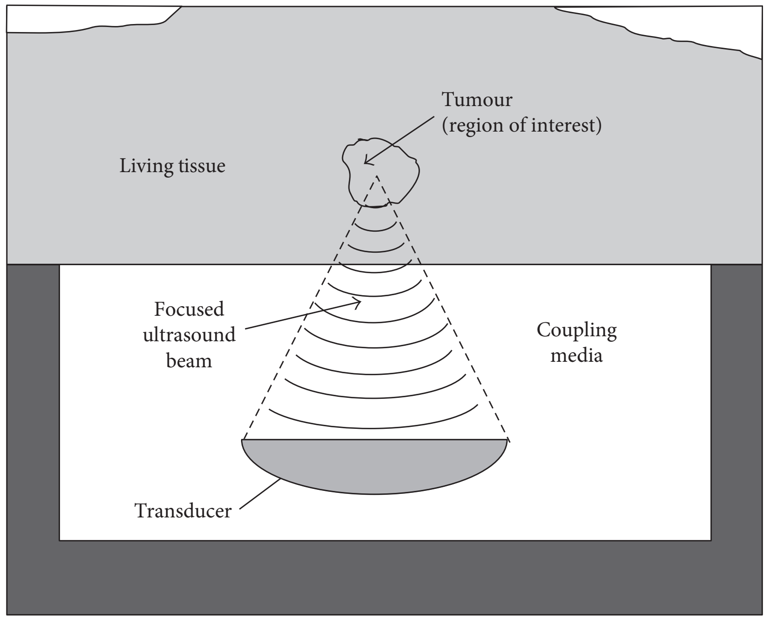


Figure 1: Objects in the HIFU therapy.

more noticeable if several images are required for the therapy planning. Different image segmentation techniques can be used for the identification of regions. The watershed transform [6] is a popular segmentation method in medical imaging [7]. Segmentation of MR images, with solutions based on the watershed segmentation algorithm, has been proposed before in other studies. In [7], an improvement to the watershed transform that enables the introduction of prior information in its calculation, in the form of markers generated from atlases, was presented. With the additional information, they limited the oversegmentation that occurs when segmenting medical images. The authors applied the algorithm to knee cartilage segmentation and white matter/gray matter segmentation in MR images, while demonstrating similar or superior performance to that of manual segmentation by experts, at an average of 0.94. To get a precise liver segmentation in abdominal MR images, Masoumi et al. [8] proposed an algorithm that utilizes MLP neural networks to extract features of the liver region to be used with the watershed algorithm. The extracted features are used to monitor the quality of the segmentation using the watershed transform and adjust the required parameters automatically. The average accuracy they achieved was 0.94 while running faster than other methods. Similar approaches have been used in [9, 10], to segment regions in brain and breast images in order to help with medical diagnosis, specifically cancer for the latter.

The fast execution of an image segmentation task has been an object of particular interest and has been addressed before in Shasidhar et al. [11] performing modifications in the standard segmentation algorithm and in Rowińska and Goćławski [12] using graphics processing units (GPUs) as an alternative to improve the execution of the segmentation algorithm. Both approaches work with the fuzzy c-means (FCM) algorithm. Approaches that consider different segmentation algorithms could be implemented.

The purpose of the present research work is to implement an efficient segmentation scheme for the MR images used for the HIFU therapy planning. In addition, it is intended that the implementation works in near real-time (i.e., that the segmented images are available after a time delay introduced by the segmentation process itself [13]) in order to address the time-consuming problem of the therapy planning. The segmentation scheme is based on the watershed method for the identification of the regions that are found on the HIFU treatment. Since the segmentation process is intended to be performed on a large amount of MR images, a thread pool was implemented in order to take advantage of all the available CPU cores for processing. This reduced the processing time of the group of MR images, but not the processing time of the watershed segmentation algorithm. The employed MR images of the present research work were obtained from a study of the distribution of heat during abscess treatment in a murine model where the transducer was positioned vis-à-vis the desired target [14].

2. Materials and Methods

An experimental protocol for the modeling of the thermal effects of the ultrasound is proposed. T1-weighted MR images were obtained from a study of the distribution of heat during abscess treatment in a murine model using a 3T MRI scanner (Achieva, Philips Healthcare). The transducer was positioned vis-à-vis the desired target. The protocol for this study was approved by the Lakehead University Animal Care Committee. The setup for the study is shown in Figure 2.

2.1. Magnetic Resonance Images. Transverse and sagittal T1-weighted MR images were obtained with a 3T MRI scanner (Achieva, Philips Healthcare). The field-of-view (FOV) is $120 \times 120 \times 48$ mm. Voxel size is 0.5 mm, and the slice thickness is 2 mm [14]. Intensity inhomogeneity is present in the

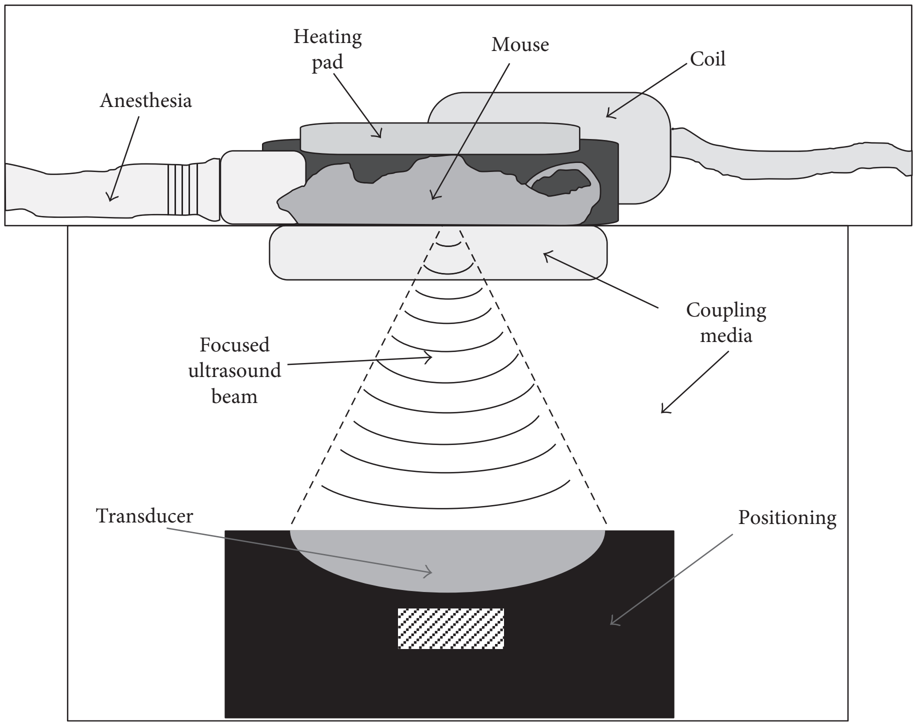


Figure 2: The experimental setup for abscess treatment in mice with focused ultrasound guided by MRI [14].

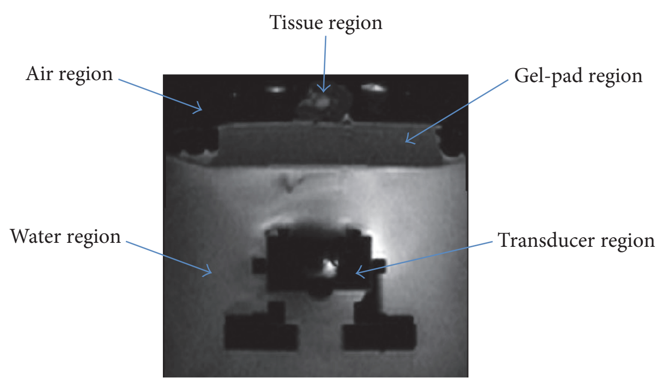


Figure 3: Main regions in the MR image (transverse image).

MR images. The regions to be identified are air, tissue, gel-pad, water, and transducer as shown in Figure 3.

2.2. Watershed Segmentation Algorithm for Image Segmentation. The watershed algorithm is based on visualizing an image in three dimensions: two spatial coordinates and an intensity value. In the algorithm, three categories of points are considered: points that belong to a regional minimum, points at which a hypothetical drop of water, if placed at the location of any of those points, would fall with certainty to a single minimum and points at which water would equally fall to more than one minimum. For a given regional minimum, the set of points satisfying the second condition is called "catchment basin." The points satisfying the third condition form crest lines on the surface and are known as "watershed lines." In watershed segmentation, a common application is the extraction of nearly uniform objects from the background [15].

2.2.1. The Use of Markers in Watershed Segmentation Algorithm. If the watershed segmentation algorithm is applied directly to the image, oversegmentation will be obtained due to irregularities in the image, yielding a useless result [7, 15]. Oversegmentation is controlled by means of markers, as a tool that brings additional knowledge to the segmentation algorithm. For the generation of markers, two main steps should be considered: preprocessing and definition of criteria that markers must fulfill [15].

For the proposed segmentation scheme, during preprocessing, noise is removed using a Gaussian filter. Then, five separate markers were defined for the watershed segmentation algorithm: three internal (tissue, gel-pad, and

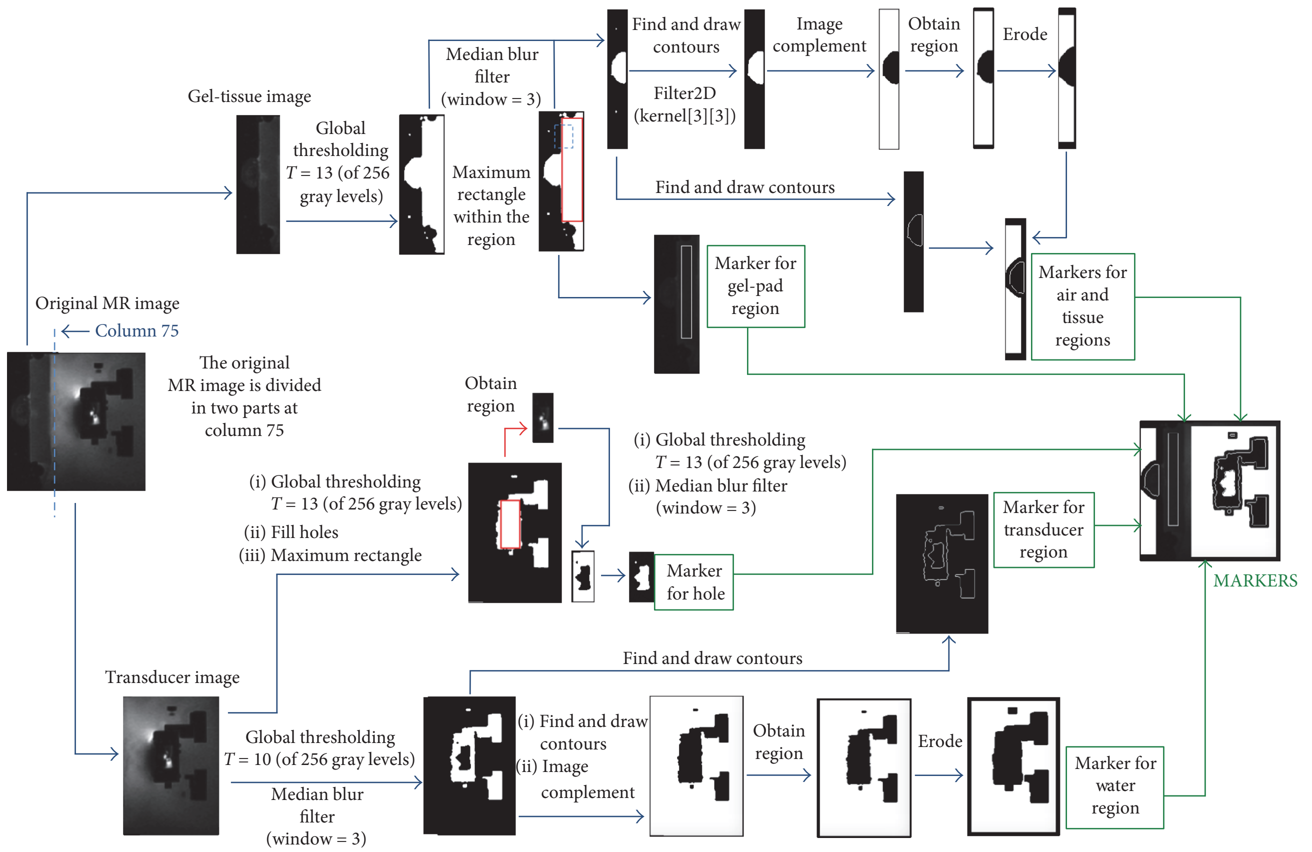


Figure 4: Definition of markers for the watershed segmentation algorithm. The image with the required markers to perform the segmentation is labeled as "MARKERS."

transducer) and two external (air and water). All the markers were stored in a separate image, each with a different value (so that the watershed algorithm could differentiate each one) and a black background.

The overall process to select the markers for the watershed segmentation algorithm is shown in Figure 4.

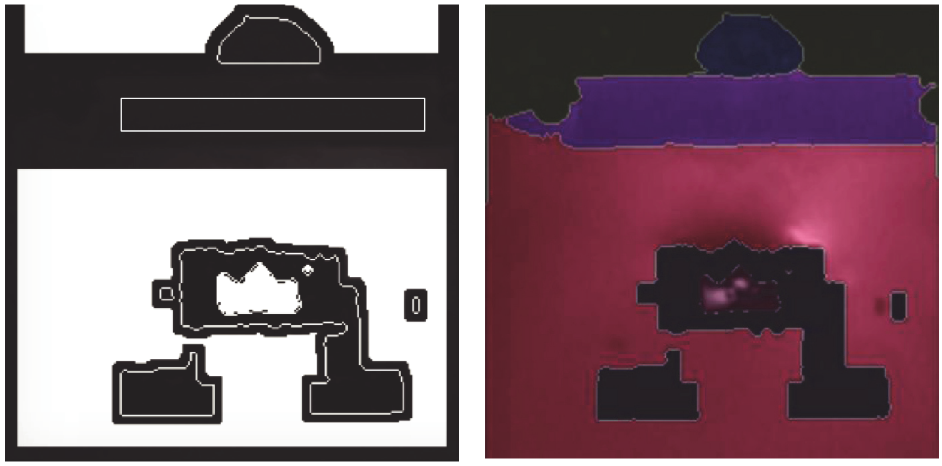
2.2.2. Definition of the Gel-Pad and Tissue Region Markers. The first step was to separate the air, gel-pad, and tissue from the transducer and the water. By reviewing the images in all of the groups, it was identified that the gel-pad and tissue are always present before the column 75 of the images and that the tissue is always smaller than the gel-pad. This information was used to create the markers for the air, tissue and gel-pad from the left image (henceforth called gel-tissue image), and the marker for the transducer and water from the right image (henceforth called transducer image).

The gel-pad region was inscribed as a rectangular region. An algorithm to find the largest rectangle submatrix on a binary matrix [16] was used on the gel-tissue image. The gel-tissue image was binarized using an intensity of 15 in a grayscale of 256 gray levels, and, since the gel-pad was larger than the tissue, the maximum rectangle found in the image was inscribed in the gel-pad.

To obtain the marker for the tissue region, the gel-tissue image was divided taking the column that corresponds to the upper left corner of the gel-pad marker as frontier. A binary matrix was obtained using a threshold of 13. A median blur filter, with a blur window of 3, was used in order to improve the shape of the tissue contours. From the resulting binary image, the contours of the tissue region were obtained with the "findContours" function in OpevCV [17]. If more than one contour was present, the one with the largest area was selected as the tissue marker.

2.2.3. Definition of the Air Region Marker. The tissue marker was used as the base for the air marker. It was filled with a different value than the air marker, and then a convolution using the kernel {{1, 1, 1}, {1, 0, 1}, and {1, 1, 1}} was performed until there were no black pixels within the tissue. The air region resulted from the negative mask of the convoluted tissue region. The resulting binary image was eroded two times in order to have a separation between the marker of the air region and the marker of the tissue region.

2.2.4. Definition of the Transducer Region Marker. A binary threshold with an intensity of 10 was applied to the transducer image that was previously obtained. Then, a median filter was used to remove any noise surrounding the transducer region. Finally, the contours of the image were found with the "findContours," "approxPolyDP," and "drawContours" functions. An additional marker was needed for the transducer object: the hole inside it. To obtain it, the largest rectangle was found within the transducer marker, and the



(a) (b)

Figure 5: Application of the watershed segmentation algorithm with markers. (a) Image containing all the generated markers. (b) Segmentation of the regions of interest with the watershed algorithm.

area inside of the rectangle was segmented and thresholded with an intensity of 13. All the black pixels were converted to the hole's value, and all white pixels were turned into black so that only the hole remained. Finally, a median filter was used to remove any pixels that were not part of the transducer hole.

2.2.5. Definition of the Water Region Marker. The transducer marker was used as the base for the water marker. The complement of the transducer marker was obtained and was consequently eroded twice so that there existed a separation of the water and transducer regions.

2.3. Processing of Image Groups. Once all the markers are generated, a marker image that contains them was created. A sample of all the markers in a single image can be seen in Figure 5(a). That marker image can be used with the watershed algorithm to segment the five regions of interest. The result of a segmentation of a sample image can be seen in Figure 5(b).

The segmentation has to be performed on a group (data set) of 224 MR images. In order to achieve a near real-time processing, a thread pool was used to apply the proposed solution to the data set in parallel.

In order to carry out the thread pool tests, the data set was segmented using the watershed segmentation algorithm (with markers). A Hewlett-Packard (HP) Z420 workstation with a 4x Intel® Xeon® CPU E5-1607 v2 @3.00GHz processor and 8GB RAM was used for the segmentation. The processor has four cores, and the operating system in this workstation is Ubuntu 14.04.1 LTS. The segmentation of the data set is done with two experiments: using one CPU core in the thread pool and using four CPU cores in the thread pool.

3. Results

F-measure was considered for the evaluation of the image segmentation quality [18]. This evaluation method has been used for the evaluation of the segmentation quality in previous research work [4].

Segmentation of the five regions of interest (air, tissue, gel-pad, transducer, and water regions) was performed with the watershed segmentation algorithm using markers over the group of MR images. The obtained results, evaluated with F-measure, are shown in the set of boxplots in Figure 6. In the figure, the air, tissue, gel-pad, transducer, and water regions are labeled as "Air," "Tissue," "Gel-Pad," "XDCR," and "Water," respectively.

When the segmentation was performed with one thread in the thread pool, the segmentation process was carried out using all the CPU cores available at different times as shown in Figure 7, resulting in an average execution time of 11.8811 sec. When the segmentation was performed with four threads in the thread pool, as shown in Figure 8, the execution time was 3.0682 sec.

4. Discussion

The use of watershed segmentation algorithm with markers yielded results that were evaluated with F-measure with medians above 0.8 for each region as shown in Figure 6.

The use of watershed segmentation algorithm with markers yielded results that were evaluated with F-measure. For the air region, the obtained medians were 0.9657 and 0.9510 in transverse and sagittal images, respectively. For the transducer regions the obtained medians were 0.9416 and 0.8821 in transverse and sagittal images, respectively. For the water region, the obtained medians were 0.9769 and 0.9459 in transverse and sagittal images, respectively. In the case of the tissue region, the obtained minimum values were 0.6219 and 0.5954 in transverse and sagittal images, respectively. Despite this fact, the results as a group yielded medians of 0.9224 and 0.9241 in transverse and sagittal images, respectively. On the other hand, for the gel-pad region, the obtained minimum value was 0.5701 in sagittal images. Despite this fact, the results as a group yielded a median of 0.9553 in this case.

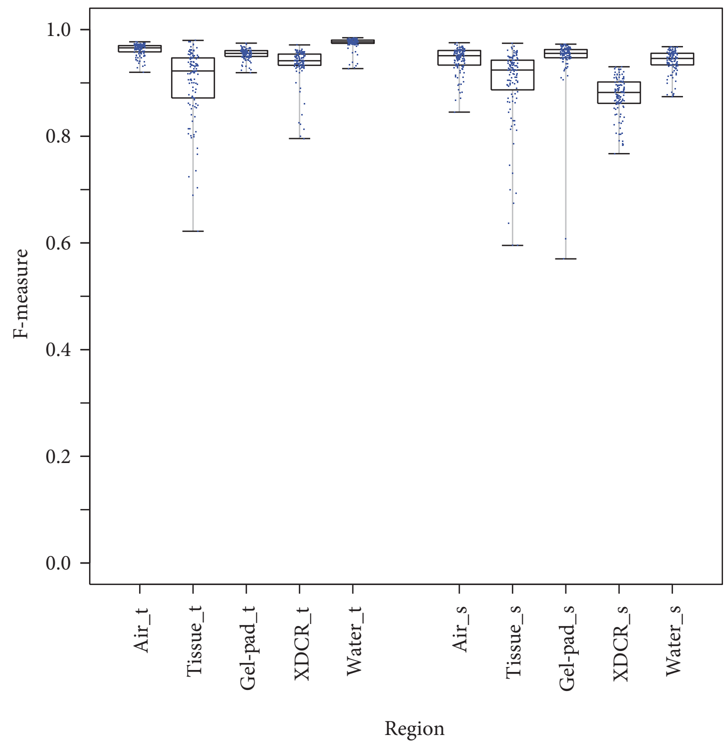


FIGURE 6: Watershed segmentation using markers in each region of interest. The letters "t" and "s" stand for transverse and sagittal planes, respectively. This graph was generated with R [19].

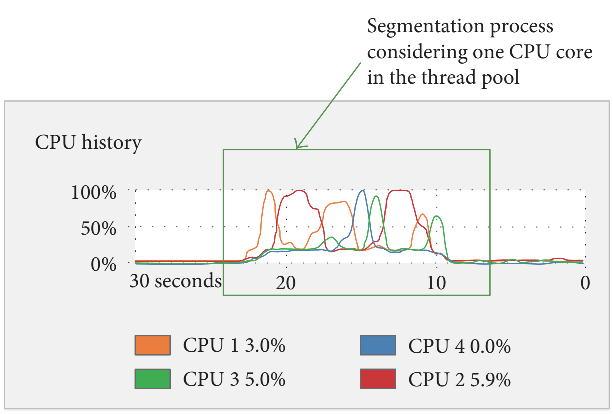


FIGURE 7: Segmentation process considering one CPU core in the thread pool displayed in the system monitor of the HP Z420 workstation. For a group of 224 images, the total execution time was 11.8811 sec.

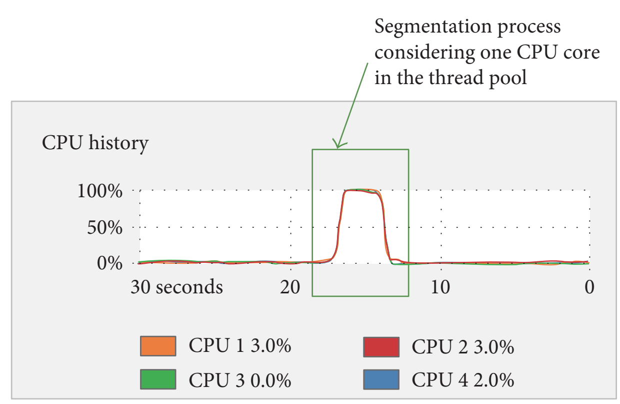


FIGURE 8: Segmentation process considering one CPU core in the thread pool displayed in the system monitor of the HP Z420 workstation. For a group of 224 images, the total execution time was 3.0682 sec.

From the thread pool results shown in Figures 7 and 8, it can be appreciated that in both cases, the CPU is using all its cores to perform a given segmentation task. However, the configuration of the thread pool considering the maximum number of cores (in this case, the use of four cores) makes the CPU work with its resources at the maximum capacity and at the same time. This results in the performance of the task in less time than using the configuration of one single core in the thread pool. An average speedup of 3.8723 in comparison with the version that only uses a single CPU core was obtained. This average speedup could be increased further if more CPU cores were available.

5. Conclusions

The implementation of the watershed segmentation algorithm with markers was carried out to address the problem of segmentation of MR images for the HIFU therapy planning. The achievement of the best possible identification of objects in the MR images was sought through the previous knowledge of the image features in order to generate proper markers for the watershed segmentation algorithm. Moreover, it was intended to achieve a time-efficient segmentation process when working with image groups.

The achievement of a functional segmentation process is feasible, but the preliminary image processing steps are required in order to define the markers for the segmentation algorithm. Despite the presence of intensity inhomogeneity in the employed MR images, the use of segmentation methods especially designed to address this problem was not necessary at all because the watershed segmentation algorithm with markers proved to have enough capacity to deal with this problem, still yielding functional segmentation results.

In order to improve the segmentation accuracy, the generation of the markers could be reevaluated. Currently, the markers are generated with information obtained from previous observation of the images. However, as can be seen in Figure 6, there were some cases in which the segmentation accuracy was really low, at around 50%. Using techniques such as neural networks [8] to generate the markers could help improve the segmentation accuracy.

By using a thread pool to apply the segmentation scheme to all the MR images of a given data set, a near real-time execution was achieved. This represents an additional contribution to solve the time-consuming problem of the HIFU therapy planning.

Another alternative that could be considered to reduce execution time is parallel processing with graphics processing units (GPUs). However, there were some limitations with the data set, primarily that the images are too small, so a redefinition of the proposed solution in order to align it with a GPU scheme is required.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

The authors wish to thank CONACYT for the support received through the scholarship no. 419184 to make this research possible.

References

[1] G. R. Ter Haar and C. Coussios, "High intensity focused ultrasound: physical principles and devices," *International Journal of Hyperthermia*, vol. 23, no. 2, pp. 89–104, 2007.

[2] K. Hynynen, "MRI-guided focused ultrasound treatments," *Science Direct Ultrasonics*, vol. 50, no. 2, pp. 221–229, 2010.

[3] K. Hynynen, "Focused ultrasound surgery guided by MRI," *Scientia Medica*, vol. 3, no. 5, pp. 62–71, 1996.

[4] D. A. Santana-Calvo, A. Vargas-Olivares, S. Pichardo, L. Curiel, and J. E. Chong-Quero, "Evaluation methods of image segmentation quality applied to magnetic resonance guided high-intensity focused ultrasound therapy," in *VI Latin American Conference in Biomedical Engineering (CLAIB 2014)*, pp. 605–608, Paraná, Argentina, 2014.

[5] F. Sannholm, *Automated Treatment Planning in Magnetic Resonance Guided High Intensity Focused Ultrasound*, Aalto University, Esbo, Finland, 2011.

[6] S. Beucher and F. Meyer, "The morphological approach to segmentation: the watershed transformation," *Optical Engineering*, vol. 34, pp. 433–481, 1993.

[7] V. Grau, A. U. J. Mewes, M. Alcañiz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 447–458, 2004.

[8] H. Masoumi, A. Behrad, M. A. Pourmina, and A. Roosta, "Automatic liver segmentation in MRI images using an iterative watershed algorithm and artificial neural network," *Biomedical Signal Processing and Control*, vol. 7, no. 5, pp. 429–437, 2012.

[9] K. Mantri and S. Kumar, "MRI image segmentation using gradient based watershed transform in level set method for a medical diagnosis system," *International Journal of Engineering Research and Applications*, vol. 4, no. 11, pp. 27–36, 2014.

[10] H. Alshanbari, S. Amain, J. Shuttelworth, K. Slman, and S. Muslam, "Automatic segmentation in breast cancer using watershed algorithm," *International Journal of Biomedical Engineering and Science*, vol. 2, no. 2, pp. 1–6, 2015.

[11] M. Shasidhar, V. Sudheer Raja, and B. Vijay Kumar, "MRI brain image segmentation using modified fuzzy C-means clustering algorithm," in *International Conference on Communication Systems and Network Technologies*, pp. 473–478, Katra, Jammu, 2011.

[12] Z. Rowińska and J. Gocławski, "CUDA based fuzzy C-means acceleration for the segmentation of images with fungus grown in foam matrices," *Image Processing & Communication*, vol. 17, no. 4, pp. 191–200, 2013.

[13] Federal Standard 1037C, "Glossary of telecommunications terms," http://www.its.bldrdoc.gov/fs-1037/fs-1037c.htm.

[14] B. Rieck, D. Bates, K. Zhang et al., "Focused ultrasound treatment of abscesses induced by methicillin resistant *Staphylococcus aureus*: feasibility study in a mouse model," *Medical Physics*, vol. 41, no. 6, article 063301, 2014.

[15] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, p. 954, Upper Saddle River, NJ, Pearson Prentice Hall, 2008.

[16] A. Agrawal, "Find largest sub-matrix with all 1s (not necessarily square)," http://tech-queries.blogspot.mx/2011/09/.

[17] "OpenCV," http://opencv.org/.

[18] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 530–549, 2004.

[19] "The R project for statistical computing," https://www.r-project.org/.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 3895164, 15 pages
https://doi.org/10.1155/2017/3895164

Research Article

A New Feature Ensemble with a Multistage Classification Scheme for Breast Cancer Diagnosis

Idil Isikli Esener,[1] Semih Ergin,[2] and Tolga Yuksel[1]

[1]*Department of Electrical Electronics Engineering, Bilecik Seyh Edebali University, 11210 Bilecik, Turkey*
[2]*Department of Electrical Electronics Engineering, Eskisehir Osmangazi University, 26480 Eskisehir, Turkey*

Correspondence should be addressed to Idil Isikli Esener; idil.isikli@bilecik.edu.tr

Received 7 January 2017; Revised 11 March 2017; Accepted 6 April 2017; Published 19 June 2017

Academic Editor: Yong Yang

A new and effective feature ensemble with a multistage classification is proposed to be implemented in a computer-aided diagnosis (CAD) system for breast cancer diagnosis. A publicly available mammogram image dataset collected during the Image Retrieval in Medical Applications (IRMA) project is utilized to verify the suggested feature ensemble and multistage classification. In achieving the CAD system, feature extraction is performed on the mammogram region of interest (ROI) images which are preprocessed by applying a histogram equalization followed by a nonlocal means filtering. The proposed feature ensemble is formed by concatenating the local configuration pattern-based, statistical, and frequency domain features. The classification process of these features is implemented in three cases: a one-stage study, a two-stage study, and a three-stage study. Eight well-known classifiers are used in all cases of this multistage classification scheme. Additionally, the results of the classifiers that provide the top three performances are combined via a majority voting technique to improve the recognition accuracy on both two- and three-stage studies. A maximum of 85.47%, 88.79%, and 93.52% classification accuracies are attained by the one-, two-, and three-stage studies, respectively. The proposed multistage classification scheme is more effective than the single-stage classification for breast cancer diagnosis.

1. Introduction

Cancer is a group of body cells that grow and proliferate abnormally and uncontrollably because of damaged DNA (deoxyribonucleic acid). This group of body cells, known as tumors, may be either benign or malignant. Benign tumors are not cancerous and life-threatening as they do not spread to other tissues or organs of the body. In stark contrast to benign tumors, malignant ones tend to be metastasized and may generally be fatal.

Breast cancer originates in a breast tissue. It is the most frequently diagnosed cancer among women, and it is 100 times more common in women than in men [1]. Worldwide, breast cancer is the second major cause of female deaths resulting from cancer [2]. There is no known way to prevent breast cancer, but mortality can be reduced with early diagnosis [3]. Radiological screening is the most important action to take for early diagnosis [4]. Although mammography is known as the most effective radiological screening technique both for breast investigation and diagnosis, the subtle difference of X-ray permeability between normal and abnormal regions makes cancer detection difficult [5]. This difficulty is aggravated as the breast tissue type becomes denser. Moreover, human factors heavily affect the interpretation of mammogram images. A computer-aided diagnosis (CAD) system detects and diagnoses cancer without these negative factors [6]. Hence, using a CAD system increases the sensitivity of cancer detection by providing radiologists a second opinion.

Classification accuracy of CAD systems is directly affected by detection of suspicious regions for breast cancer, namely region of interest (ROI), from whole-breast mammogram images. Besides the low-contrast problem, the digitization noise in mammograms also affects the success of ROI detection negatively; and noise reduction is required to improve the image quality [7, 8]. Hence, preprocessing is necessary and should be the first of the four stages in a

CAD system. Some studies have tried to overcome the problem of low contrast using histogram processing operations [8–11], morphological operations [12], and statistics theory [13], while unsharp filtering [8], wavelet transform [12, 13], and median filtering [14, 15] are the most common noise reduction.

In the second stage of a CAD system, the ROIs are detected from entire breast images. ROI detection in the past decade was generally performed using wavelet transforms [16], segmentation algorithms [17, 18], and edge operators [19].

The efficiency of a CAD system is directly related to the efficacy of data representation. Feature extraction, which is the third stage, is an undoubtedly important task for pattern recognition and is implemented with a remarkable number of techniques in several studies. Specifically, there are statistical techniques [20–25], model-based techniques [26, 27], graph-theoretic approaches [28], and signal processing techniques that compute breast tissue features from pixel characteristics [22, 29] or frequency spectrum [21, 23, 24, 30–32] for breast cancer diagnosis on a mammographic image. Additionally, there are various studies using mammographic features [20, 33–36] like shape, spicule index, contour, size, density, and brightness.

Finally, in the fourth stage, extracted discriminative features are used for the classification of ROIs into normal, benign, and malignant lesions. Artificial neural networks [20, 27, 29, 36, 37], support vector machines (SVMs) [20, 25, 30, 31, 33, 38, 39], subspace learning algorithms [22, 25, 38, 39], Bayes, decision tree, and k-nearest neighbor classifiers [20, 25] are well-known classifiers used for mammogram classification.

Mammogram-based breast cancer diagnosis studies can be categorized as microcalcification detection, mass detection, and mass recognition. Pal et al. presented a multistage system for microcalcification detection [27]. This multistage system first classifies a mammogram image as normal or abnormal; then, for an abnormal image, it detects the regions with microcalcification. The authors extracted statistical features on manually detected ROIs and implemented feature selection and classification using a multilayer perceptron neural network [27]. Lado et al. developed an extended generalized additive model (GAM) involving interaction of breast tissue factors to reduce the false-positive rate for microcalcification detection [16]. The authors stated that the false-positive rate has decreased to 0.74 per image from 1.46 when the breast tissue type is integrated into the GAM. Similarly, Malar et al. studied the effectiveness of breast tissue type integration on microcalcification detection using an extreme learning machine and achieved an accuracy of 94% using wavelet-based features [40]. Since the number of cells with microcalcification is smaller than the number of healthy cells, microcalcification detection is an unbalanced classification problem. Bria et al. proposed a cascaded five-classifier approach to eliminate the predominance of healthy cells [41]. In this approach, the first classifier initially discriminates the normal and abnormal cells and later benign microcalcification clusters (μCs) and false detections of normal cells are eliminated using a RankBoost classifier. The resultant malignant μCs are evaluated by the next classifier, and the process goes on until the μCs from the last classifier are obtained. Ultimately, final μCs are selected according to their probability maps with 93% accuracy. Kekre et al. [17, 18] segmented mammogram images using a vector quantization technique for mass detection. They computed the areas of each region on the segmented images and classified the region having the largest area as a mass. Hachama et al. used an image registration for mass detection [42]. Savitha et al. suggested that analyzing mammogram images in the complex plane will increase the accuracy of mass detection [43]. They mapped the mammogram images into a complex plane and classified them using a fully complex-valued relaxation neural network with an accuracy of 97.84%. Vallez et al. stated that lesion detection and recognition accuracy can be increased by using predefined breast tissue type information [25]. The classification accuracy rate has been increased to 91% from 78% in their study. Guliato et al. suggested that the previously proposed polygonal modeling [44] is an effective method for mammogram classification as it helps in noise reduction while preserving the important features [45]. Oliver et al. proposed a knowledge-based approach for the automatic detection of microcalcifications and clusters in mammographic images [37]. In this approach, local features that characterize the morphology of microcalcifications are first extracted to create a dictionary of visual words by a bank of filters. Then, feature selection is accomplished by using a boosted classifier for microcalcification detection. Finally, the cluster detection is achieved at 80% sensitivity by locally integrating the individual microcalcification probability images.

In this paper, a new and effective feature ensemble with a multistage classification is proposed to be implemented in a CAD system for breast cancer diagnosis. The result is verified using a publicly available mammogram image dataset collected during the Image Retrieval in Medical Applications (IRMA) project. For the preprocessing stages, contrast enhancement and noise reduction operations are first executed on each mammogram ROI in the database by applying a histogram equalization followed by a nonlocal means (NLM) filtering [46]. The local configuration pattern (LCP) algorithm [47] is then applied to obtain LCP-based feature vectors from the mammographic images. Then, some statistical and frequency-domain features are extracted and concatenated with the LCP-based feature vectors. Eventually, these feature vectors formed by LCP-based, statistical, and frequency-domain features are classified as normal, benign, and malignant using eight different popular classifiers via cross validation. The classification process is performed in three different cases in this study. In the first case, called a one-stage study, the feature vectors are directly classified into three classes. In the second case, called a two-stage study, the feature vectors are initially categorized according to their breast tissue types, and are subsequently classified as normal, benign, and malignant. In the third case, called a three-stage study, the feature vectors are first classified according to their breast tissue types. Afterward, they are classified as normal and abnormal. At the third stage of this case, the feature vectors labeled as abnormal classes are categorized as benign and malignant. Moreover, a classifier combination via a majority voting of the most successful three classifiers is employed for both the two- and three-stage studies.

This paper is organized as follows. The preprocessing and the whole feature extraction procedure realized in this paper are explicated and all of the classification methods and the evaluation metrics are briefly described in the following section. Discussions on the experimental studies and the obtained results are given in Section 3, whereas the main conclusions are precisely specified in the last section.

2. Materials and Methods

2.1. Database. It is very important to work on images with their ground truths for medical imaging applications [48]. In this study, a publicly available mammogram dataset constructed during the IRMA project is used [49]. This dataset consists of 12 classes defined by the Breast Imaging Reporting and Data System (BI-RADS). There are four breast tissue classes (fatty, fibroglandular, heterogeneously dense, and extremely dense) and three health status classes (normal, benign tumor, and malignant tumor) for each breast tissue type. There are 233 mammogram ROI parts, lower-dimensional mammogram images that consist of just healthy/cancerous regions of the whole breast, for each class, and therefore, a total of 2796 parts are available in the dataset [49]. The ROI parts of each class are classified using cross-validation technique. It implicitly means that 210 of 233 parts (90%) in each class are used for training while the remaining 23 of 233 parts (10%) are treated as the test parts. The process is repeated for each fold in the cross-validation technique, and the average classification accuracy for each classifier is obtained.

2.2. Preprocessing. In the preprocessing stage, a histogram equalization followed by the NLM filtering is applied on the mammogram parts [48]. The NLM filter is an adaptive smoothing filter that changes the window size according to the similarity between neighborhoods of any two pixels as well as preserves the fine details by computing a weighting function according to the derivatives in the corresponding search window [46, 48]. Given a discrete noisy image $v = \{v(i) | i \in I\}$, the filtered value $\mathrm{NL}[v(i)]$ of any pixel is computed as

$$\mathrm{NL}[v(i)] = \sum_{j \in I} w(i,j) \cdot v(j), \tag{1}$$

where $w(i,j)$ refers to the weight coefficient computed utilizing the similarity between pixels i and j and satisfies the conditions $0 \le w(i,j) \le 1$ and $\sum_j w(i,j) = 1$.

The similarity between pixels i and j is measured as the Gaussian weighted Euclidean distance, $\|v(N_i) - v(N_j)\|^2_{2,\sigma}$, where σ $(\sigma > 0)$ is the standard deviation of the Gaussian kernel, whereas $v(N_i)$ and $v(N_j)$ are the neighborhoods of pixels i and j in the similarity window [48]. The pixels with larger weights indicate a similar neighborhood as it can be understood by analyzing (2). Z_i and h in (2) refer to the normalizing constant and the degree of filtering, respectively.

$$w(i,j) = \frac{1}{Z_i} \cdot e^{-\|v(N_i) - v(N_j)\|^2_{2,\sigma}/h^2}, \tag{2}$$

$$Z_i = \sum_j e^{-\|v(N_i) - v(N_j)\|^2_{2,\sigma}/h^2}. \tag{3}$$

2.3. Feature Extraction. The most essential stage in CAD systems, as well as in any pattern recognition problem, is the feature extraction in which data is represented in a low-dimensional space by the most descriptive features that maximize and characterize the interclass differences. In this study, three groups of features are concatenated to construct the feature vectors. The first group is LCP-based features obtained using LCP algorithm, while the second and third groups are some statistical and frequency-domain features, respectively.

2.3.1. Local Configuration Pattern. The local binary pattern (LBP) is generally used for face representation and recognition in the past two decades [50–52], and it is a grayscale and rotation-invariant feature extraction technique presented by Ojala et al. [53].

The grayscale-independent LBP representation of an image I is obtained by thresholding P neighbors in the circular neighborhood of radius R with the intensity value of the central pixel as given in (4).

$$\mathrm{LBP}(P,R) = \sum_{i=0}^{P-1} u(g_i - g_c) \cdot 2^i, \tag{4}$$

$$u(x) = \begin{cases} 1, & x \ge 0 \\ 0, & x < 0. \end{cases} \tag{5}$$

The terms g_i and g_c in (4) denote the intensity values of the neighboring pixel i and central pixel c, respectively. The rotation-invariant LBP-based feature vectors are described by the idea of rotating each bit pattern circularly to a minimum value ending up with the maximum value as the last element of the feature vectors. Equation (6) introduces the mathematical representation of this idea where the term $\mathrm{LBP}^{\mathrm{riu2}}$ refers to the rotation-invariant LBP-based feature vectors.

$$\mathrm{LBP}^{\mathrm{riu2}}(P,R) = \begin{cases} \sum_{i=0}^{P-1} u(g_i - g_c), & U(\mathrm{LBP}(P,R)) \le 2 \\ P+1, & \text{otherwise}. \end{cases} \tag{6}$$

The quantization of gray-level differences to binary levels sometimes causes undesirably the same LBP representations although the neighborhoods are relatively different. This problem is solved by computing the local variance (VAR) of each pattern, and the joint histogram (O) is formed. μ in (7) refers to the average intensity of the neighboring pixels.

$$\mathrm{VAR} = \frac{1}{P} \sum_{i=0}^{P-1} (g_i - \mu)^2, \tag{7}$$

$$O = \frac{\mathrm{LBP}^{\mathrm{riu2}}}{\mathrm{VAR}}. \tag{8}$$

The LBP algorithm is stated to be an effective technique for detecting local structures; however, the $\mathrm{LBP}^{\mathrm{riu2}}$ feature vectors for patterns having equal variances may be the same although they have different configurations [47]. Guo et al. proposed a microscopic (MiC) descriptor that defines the microscopic configuration of an image by a linear configuration model as a solution to this problem [47]. In this model, the optimal weights (A_{L}) of the neighboring pixels are calculated via the least square estimation technique to form the central pixel. For the conservation of being a rotationally invariant characteristic, a one-dimensional Fourier transform of optimal weight vectors is computed and H_{L} values are obtained. The magnitude of H_{L} is defined as the MiC feature of a pattern.

The local configuration pattern (LCP) is a technique that describes the local structures and microscopic configuration of a pattern together, where the LCP-based feature vector of an image is obtained by concatenating the microscopic configuration of each pattern in an image with their joint histogram as [24]

$$\mathrm{LCP} = \left[[H_0|; O_0]; [H_1|; O_1]; \ldots; \left[H_{q-1}|; O_{q-1}\right] \right], \tag{9}$$

where q is the number of patterns in an image.

2.3.2. Statistical Features. Some significant and descriptive statistical features of each LCP-based feature vector are calculated as the second group of features to increase the data representability of the feature vectors. Energy is one of the most important statistical features of any distribution, and hence, the energy values of LCP-based feature vectors are evaluated. The mean, maximum, minimum, and mean energy of each LCP-based feature vector are additionally computed as statistical features. In the statistical theory, the variance, skewness, and kurtosis are defined as variation criterions. Owing to the large variations between healthy and cancerous regions on a mammogram image, these criterions are also calculated. Moreover, the standard deviation, energy variance, and area descriptor [54] of LCP-based feature vectors are additional variation-related features used in this study. Radiologists state that cancerous regions and malignant regions have more irregular distribution than healthy regions and benign regions, respectively. This statement corresponds to entropy in statistics. Therefore, the entropy of each LCP-based feature vector is calculated to measure this irregularity as a feature. The statistical features utilized in this study and their mathematical representations for the $N \times 1$ dimensional feature vectors are listed in Table 1.

Table 1: Statistical features and their mathematical representations.

Energy	$\sum_{i=1}^{N} X_i^2$
Mean	$\mu = \frac{1}{N} \cdot \sum_{i=1}^{N} X_i$
Variance	$\mathrm{Var} = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (X_i - \mu)^2$
Maximum	$\mathrm{Maximum}\{X_i \mid i = 1, 2, \ldots, N\}$
Minimum	$\mathrm{Minimum}\{X_i \mid i = 1, 2, \ldots, N\}$
Standard deviation	$\sigma = \sqrt{\mathrm{var}}$
Skewness	$\frac{1}{\sigma^3} \cdot \sum_{i=1}^{N} (X_i - \mu)^3$
Kurtosis	$\frac{1}{\sigma^4} \cdot \sum_{i=1}^{N} (X_i - \mu)^4$
Area descriptor [50]	$\frac{\sigma}{\mu}$
Mean energy	$\mu_{\mathrm{Energy}} = \frac{1}{N} \cdot \sum_{i=1}^{N} X_i^2$
Energy variance	$\frac{1}{N-1} \cdot \sum_{i=1}^{N} (X_i^2 - \mu_{\mathrm{Energy}})^2$
Entropy	$-\sum_{i=1}^{N} p(X_i) \cdot \log_2 p(X_i)$

2.3.3. Frequency-Domain Features. The third group of features computed in this study is the frequency-domain features. Frequency-domain features are determined by applying a two-level two-dimensional discrete wavelet transform (2D-DWT) using Daubechies1 (db1) wavelet function on the preprocessed mammogram images, and finally, 16 sub-bands for each mammogram image are obtained. The energy values of each sub-band are computed since the brightness is one of the most significant issues for breast cancer diagnosis. db1 function is a type wavelet in wavelet analysis. The mother function $\psi(t)$ of db1 wavelet is described as [55]

$$\psi(t) = \begin{cases} 1, & 0 \le t < \frac{1}{2} \\ 1, & \frac{1}{2} \le t < 1 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

2.3.4. Feature Vector Construction. The preprocessed mammogram parts are decomposed into four sub-bands that are LL (low-low), LH (low-high), HL (high-low), and HH (high-high) by a one-level 2D-DWT utilizing the db1 wavelet. Several parameter values are experienced in the LCP transform, and ultimately, the LCP algorithm is applied on each sub-band using 8 neighbors in the circular neighborhood of radius 2. Therefore, 81×1 dimensional LCP vectors of each sub-band are constructed. The endmost values in those LCP vectors are appreciably high; therefore, they are removed to get rid of their domination over other features. The remaining 80-dimensional feature vectors of each sub-band {LL-LH-HL-HH} are then weighted with the respective coefficients {1.4-1-1-0} concluded as the most efficient coefficients by [5]. Then, they are summed

Table 2: Feature vector construction process.

108-dimensional feature vector content		
LCP-based feature vector	Statistical features	Frequency-domain features
80×1	12×1	16×1
The 80-dimensional LCP-based feature vector		LLLL: energy
		LLLH: energy
	LCP: energy	LLHL: energy
	LCP: mean	LLHH: energy
	LCP: variance	LHLL: energy
	LCP: maximum	LHLH: energy
	LCP: minimum	LHHL: energy
	LCP: standard deviation	LHHH: energy
	LCP: skewness	HLLL: energy
	LCP: kurtosis	HLLH: energy
	LCP: area descriptor	HLHL: energy
	LCP: mean energy	HLHH: energy
	LCP: energy variance	HHLL: energy
	LCP: entropy	HHLH: energy
		HHHL: energy
		HHHH: energy

up to form an 80-dimensional feature vector for each mammogram part [48].

In order to increase the representative power of the feature vectors, 12 statistical features computed from the LCP-based feature vectors, and 16 frequency-domain features evaluated from the sub-bands obtained by the decomposition of the preprocessed mammogram ROI parts using the two-level 2D-DWT are concatenated to the LCP-based feature vectors [48]. Consequently, 108-dimensional feature vectors are extracted from each ROI part. The statistical features are extracted from the LCP-based feature vectors instead of extracting them directly from the mammogram texture to amplify the discriminative power of the LCP-based feature vectors. The frequency-domain features, which are the energy values of each sub-band in the spatial domain, are extracted since the brightness is one of the most significant issues for breast cancer diagnosis, and changes in the brightness in a mammogram image are clearly observed in the spatial frequency. Table 2 summarizes the feature vector construction process. In Table 2, the phrase "LCP: energy" refers to the energy value of an LCP vector whereas "LLLL: energy" is the energy of the LLLL (low-low-low-low) sub-band.

2.4. Classifiers

2.4.1. Fisher's Linear Discriminant Analysis. Fisher's linear discriminant analysis (FLDA) tries to find a projection matrix that projects the training data onto a low-dimensional space that maximizes between-class variance as well as minimizing within-class variance [48, 56]. This is known as the Fisher maximization criterion and is defined as

$$J(\vec{w}) = \frac{\vec{w}^T \cdot S_B \cdot \vec{w}}{\vec{w}^T \cdot S_W \cdot \vec{w}}, \tag{11}$$

where $\vec{w}$, S_B, and S_W refer to the projection vectors and between-class and within-class scatter matrices, respectively.

On the test stage of FLDA, any test vector is projected via $\vec{w}$ projection vectors, and distances to the training vectors on the low-dimensional space are calculated [48]. The decision criterion for FLDA is given as

$$K = \underset{1 \le c \le S}{\arg\min} \left\{ \left\| \vec{\Omega}^c - \vec{\Omega}_{\text{test}} \right\| \right\}, \tag{12}$$

where c is the class index, S is the total number of classes, and $\vec{\Omega}^c$ and $\vec{\Omega}_{\text{test}}$ are the projected training vector of the cth class and the projected test vector, respectively [48].

2.4.2. Linear Discriminant Classifier. Linear discriminant classifier (LDC) tries to find the weight vectors $\vec{w}$ of a linear hyperplane $g(\vec{x})$ that separates given classes [57]. The weight vectors of this hyperplane are defined by a linear combination of training feature vectors $(\vec{x})$ of each class. The linear hyperplane is characterized by the weight vectors and a threshold w_0 as

$$g(\vec{x}) = \vec{w}^T \cdot \vec{x} + w_0. \tag{13}$$

The LDC assigns any test vector $(\vec{x}_{\text{test}})$ to a class according to the sign of the projection function given in (14) for a two-class problem. The terms w_1 and w_2 in (14) refer to the class labels.

$$\vec{x}_{\text{test}} \in \begin{cases} w_1, & \vec{w}^T \cdot \vec{x}_{\text{test}} + w_0 > 0 \\ w_2, & \vec{w}^T \cdot \vec{x}_{\text{test}} + w_0 < 0. \end{cases} \tag{14}$$

2.4.3. Support Vector Machines. Support vector machines (SVMs), also known as maximum margin classifiers, determine the optimal hyperplane that maximizes the distance between the hyperplane and support vectors [58]. Support vectors are the training vectors that are nearest from each class to the hyperplane [59]. As it can classify linearly separable data, SVM can classify nonlinear data by transforming the data to a higher-dimensional space by using an appropriate kernel function [49]. If the training set is $\text{TS} = \{(\vec{x}_1, L_1), (\vec{x}_2, L_2), \ldots, (\vec{x}_M, L_M)\}$ for a two-class problem, where $\vec{x}_i$ $(i = 1, 2, \ldots, M)$ is the training data and L_i $(L_i \in \{-1, 1\})$ is the class label, the test vector is classified according to the sign of the function given as

$$f(\vec{x}_{\text{test}}) = \sum \left\{ \alpha_i \cdot L_i \cdot \left(\vec{x}_i^T \cdot \vec{x}_{\text{test}} \right) + b \right\}, \tag{15}$$

where α_i $(i = 1, 2, \ldots, M)$ are the nonzero quadratic coefficients and $(|b|/\|\vec{w}\|)$ is the perpendicular distance between the hyperplane and the origin, whereas $\vec{w}$ is the normal vector of the separating hyperplane [48].

2.4.4. Logistic Linear Classifier. The logistic linear classifier (LLC) states that a linear hyperplane can be characterized by the relationship between the dependent and independent variables of training feature vectors ($\vec{x}$) [60]. In LLC, this relationship is determined using a logistic regression analysis by computing class-conditional probability density functions of $\vec{x}$ vectors. The LLC model for a two-class problem is given by (16) where $p(\vec{x}|w_i)$, $\vec{\beta}$, and β_0 are the class-conditional probability density functions of $\vec{x}$, weight vectors for the linear hyperplane, and a threshold value, respectively.

$$\log\left(\frac{p(\vec{x}|w_1)}{p(\vec{x}|w_2)}\right) = \vec{\beta}^T \cdot \vec{x} + \beta_0. \tag{16}$$

The LLC assumes that log-linear models can be formed between classes with equal prior probabilities and covariance matrices. This assumption is equivalent to

$$p(w_1|\vec{x}) = \frac{\exp\left(\vec{\beta}^T \cdot \vec{x} + \beta_0'\right)}{1 + \exp\left(\vec{\beta}^T \cdot \vec{x} + \beta_0'\right)},$$
$$p(w_2|\vec{x}) = \frac{1}{1 + \exp\left(\vec{\beta}^T \cdot \vec{x} + \beta_0'\right)}, \tag{17}$$

$$\beta_0' = \beta_0 + \log\left(\frac{p(w_1)}{p(w_2)}\right), \tag{18}$$

where $p(w_i|\vec{x})$ and $p(w_i)$ are the probabilities of class w_i given $\vec{x}$ and prior probability of class w_i, respectively. The decision criterion for LLC is given in

$$\vec{x} \in \begin{cases} w_1, & \dfrac{p(w_1|\vec{x})}{p(w_2|\vec{x})} > 1 \\ w_2, & \dfrac{p(w_1|\vec{x})}{p(w_2|\vec{x})} < 1, \end{cases} \tag{19}$$

$$\vec{x} \in \begin{cases} w_1, & \vec{\beta}^T \cdot \vec{x} + \beta_0' > 0 \\ w_2, & \vec{\beta}^T \cdot \vec{x} + \beta_0' < 0. \end{cases} \tag{20}$$

2.4.5. Decision Tree. The principle of the decision tree classifier is to cluster any data into subgroups until all elements of any subgroup have the same class label [48, 61]. Classification rules are defined by clustering the data into the leaves, class labels, in the training stage while those rules are applied to any test sample and the leaf that the test sample reaches provides the class label of the test sample in the test stage.

2.4.6. Random Forest. The random forest classifier is an ensemble of decision tree classifiers developed to improve the classification accuracy [62]. Each tree classifier in this ensemble votes for the best class of any sample, and the resultant class label is then specified via a majority voting technique.

2.4.7. Naïve Bayes. Bayesian classifiers compute the probability of each class given any test vector ($\vec{x}$) and assign it to the class with the highest conditional probability [63]. The Bayesian decision criterion for a two-class problem is

$$P[w_1|\vec{x}] > P[w_2|\vec{x}] \Leftrightarrow \vec{x} \in w_1. \tag{21}$$

The terms $P[w_1|\vec{x}]$ and $P[w_2|\vec{x}]$ denote the posterior probabilities of classes w_1 and w_2 given $x^{\rightarrow}$, respectively, where $P(w_i|\vec{x})$ is computed as

$$P(w_i|\vec{x}) = \frac{p(\vec{x}|w_i) \cdot P(w_i)}{p(\vec{x})}, \tag{22}$$

$$p(\vec{x}) = \sum_{i=1}^{2} p(\vec{x}|w_i) \cdot P(w_i). \tag{23}$$

The terms $P(w_i)$, $p(\vec{x}|w_i)$, and $p(\vec{x})$ refer to the prior probability of class w_i, the probability of $\vec{x}$ given class w_i, and the probability density function of $\vec{x}$, respectively. One-dimensional and l-dimensional case computations of $p(\vec{x}|w_i)$ are given in (24) and (25), respectively. μ, σ, and $\sum$ in these equations are the mean, variance, and covariance matrix of the feature vectors, respectively.

$$p(\vec{x}|w_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left(-\frac{(\vec{x} - \mu)^2}{2\sigma^2}\right), \tag{24}$$

$$p(\vec{x}|w_i) = \frac{1}{(2\pi)^{1/2}\left|\sum\right|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\vec{x} - \mu)^T \cdot \sum^{-1} \cdot (\vec{x} - \mu)\right). \tag{25}$$

Naïve Bayes classifiers assume that all feature vectors are statistically independent and classify any test vector according to the Bayesian decision criterion given in (21) [63]. In this classification scheme, the probability density function for the l-dimensional case is computed as

$$p(\vec{x}) = \prod_{i=1}^{l} p(x(i)). \tag{26}$$

2.4.8. k-Nearest Neighbors. The k-nearest neighbor (kNN) classifier assigns any test vector to the respective class that its k-nearest neighbors belong at most, considering the distances between the test and training vectors in the feature space [64]. Although it is obvious that classification performance is directly related to the parameter k, there is no obvious information on the selection of k except that it should be positive and not a multiple of the total number of classes [48].

2.5. Evaluation Metrics. The metrics sensitivity (SNS), specificity (SPC), positive predictive value (PPV), negative predictive value (NPV), false-positive rate (FPR), false-negative rate (FNR), false discovery rate (FDR), false omission rate (FOR), and accuracy (ACC) are used for the evaluation of the performance of the CAD system in this study. The mathematical representations of these metrics are given in Table 3.

3. Results and Discussion

In this study, a CAD system for breast cancer diagnosis based on a multistage classification using a novel feature ensemble

Table 3: Evaluation metrics and their mathematical representations.

TP: true positive FP: false positive	TN: true negative FN: false negative
Sensitivity (SNS)	$\%\mathrm{SNS} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \cdot 100$
Specificity (SPC)	$\%\mathrm{SPC} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \cdot 100$
Positive predictive value (PPV)	$\%\mathrm{PPV} = \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \cdot 100$
Negative predictive value (NPV)	$\%\mathrm{NPV} = \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} \cdot 100$
False-positive rate (FPR)	$\%\mathrm{FPR} = \dfrac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \cdot 100$
False-negative rate (FNR)	$\%\mathrm{FNR} = \dfrac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \cdot 100$
False-discovery rate (FDR)	$\%\mathrm{FDR} = \dfrac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}} \cdot 100$
False omission rate (FOR)	$\%\mathrm{FOR} = \dfrac{\mathrm{FN}}{\mathrm{TN}+\mathrm{FN}} \cdot 100$
Accuracy (ACC)	$\%\mathrm{ACC} = \dfrac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \cdot 100$

is proposed. The feature extraction stage is achieved on mammogram ROIs that are preprocessed by applying a histogram equalization followed by the NLM filtering. The proposed feature ensemble is formed by concatenating the LCP-based, statistical, and frequency-domain features. The classification process of these features is implemented in three different cases: one-stage study, two-stage study, and three-stage study. The mammogram ROIs are classified into three classes (normal, benign, and malignant) regardless of their breast tissue types in the one-stage study while the two- and three-stage studies consider breast tissue information and make a health status classification as explicitly explained in the related subsections. Eight well-known classifiers (FLDA, LDC, linear SVM, LLC, decision tree, random forest, naïve Bayes, and kNN) are used in all of the classification cases. Additionally, the results of classifiers that show the top three performances are combined via a majority voting technique in order to improve the recognition accuracy for the both two- and three-stage studies. The block diagram of the proposed system is given in Figure 1.

3.1. Results

3.1.1. One-Stage Study. In this case of the classification scheme, the feature vectors are directly classified into three classes (normal, benign, and malignant) regardless of the breast tissue types of the mammogram images. The flowchart for the one-stage study is shown in Figure 2. The average classification accuracies and standard deviations of the classifiers for the one-stage study obtained by elevenfold cross-validation technique are shown in Figure 3. In this figure, "SVM ('p', 1)" is the SVM classifier using a linear kernel. The LLC classifier has the highest recognition accuracy (85.47%) among all classifiers. It assumes that logistic linear models can be formed between classes with equal prior probabilities. Hence, it is more applicable for the one-stage study than the other classifiers as the prior probabilities of each class in this case are equal.

The total confusion matrix of the LLC classifier obtained by elevenfold cross-validation for the one-stage study is given in Table 4. It shows that benign and malignant mammograms are distinguishable from each other. The false recognitions are caused by the confusion of the benign and malignant mammograms with the normal mammograms.

The evaluation metrics of each classifier evaluated by elevenfold cross-validation for the one-stage study are given in Table 5.

The one-stage study is also achieved using three additional sets of feature vectors in order to demonstrate the discriminative power of the proposed 108-dimensional feature vector ensemble. These sets consist of 12-dimensional statistical feature vectors, 80-dimensional LCP-based feature vectors, and 92-dimensional feature vectors concatenated by the LCP-based with statistical features. The average classification accuracies of the classifiers for the one-stage study obtained by elevenfold cross-validation technique using different feature vector sets are shown in Figure 4. It can be inferred from Figure 4 that classification accuracies are increased when 92-dimensional feature vectors are used rather than only statistical or only LCP-based features. Furthermore, 108-dimensional feature vectors provide higher recognition accuracies than the 92-dimensional feature vectors. These results obviously prove the effectiveness of the proposed feature ensemble.

3.1.2. Two-Stage Study. The recognition accuracy for breast cancer diagnosis is expected to be enhanced by the two-stage study, which is composed of the breast tissue and health status classification. In the first stage of this study, the feature vectors are classified into breast tissue classes (fatty, fibroglandular, heterogeneously dense, and extremely dense). Then, the breast-tissue-type-defined feature vectors are classified into normal, benign, and malignant classes in the second stage. The flowchart for the two-stage study is shown in Figure 5.

The average classification accuracies and standard deviations of classifiers obtained by elevenfold cross-validation technique for the two-stage study are shown in Figure 6. A maximum of 87.51% accuracy rate is attained using the FLDA classifier among eight well-known classifiers. For this case, the LLC classifier performs worse than FLDA classifier as the prior probabilities of the classes are no longer equal.

As it can be explicitly inferred from Figure 6, the top three classifiers based on performance are the FLDA, LLC, and LDC. The results of these classifiers are combined via a majority voting technique to increase the classification accuracy to 88.79%.

The total confusion matrices of the (a) FLDA, (b) LLC, and (c) LDC classifiers obtained by elevenfold cross-validation for the two-stage study and the total confusion matrix of the classifier combination obtained by elevenfold

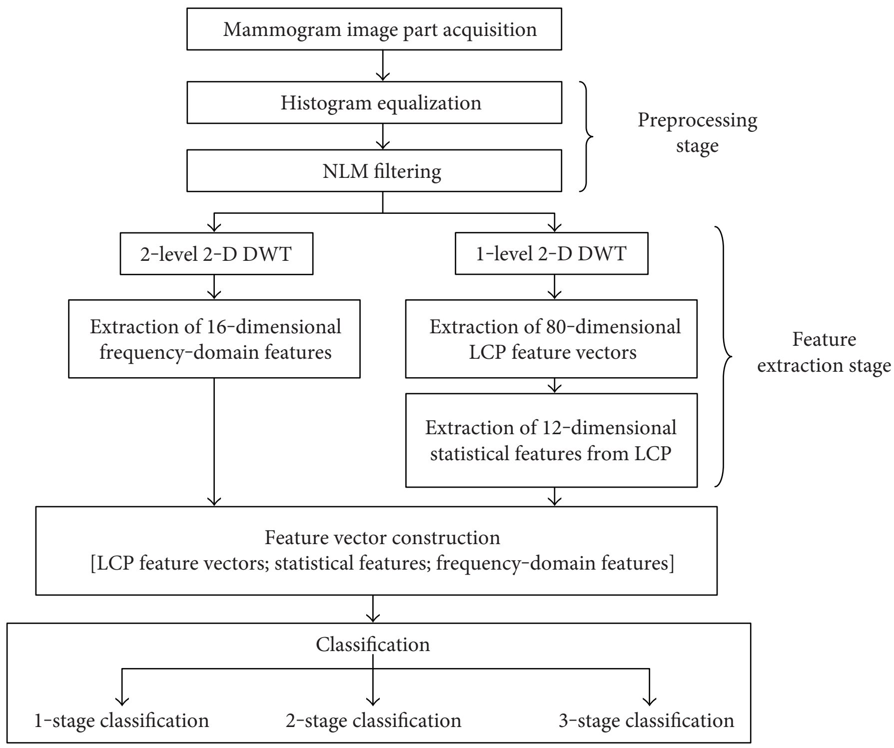


Figure 1: Block diagram of the proposed system.

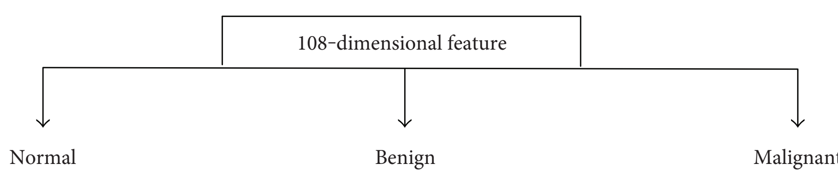


Figure 2: Flowchart designed for the one-stage study.

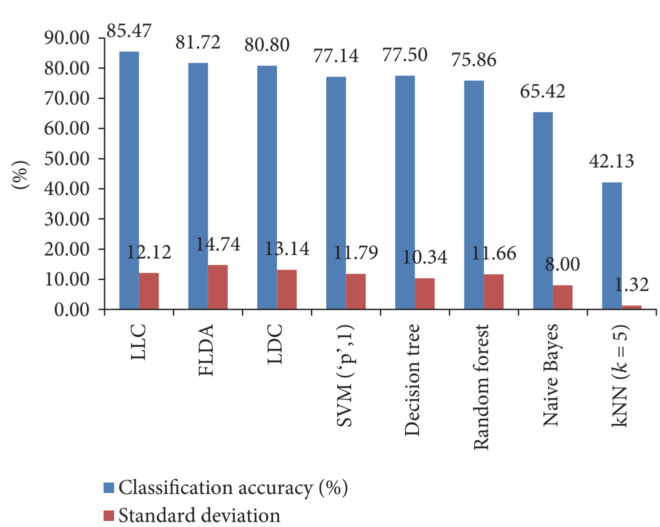


Figure 3: Average classification accuracies and standard deviations of eight different classifiers obtained by elevenfold cross-validation for the one-stage study.

Table 4: Total confusion matrix of the LLC classifier obtained by elevenfold cross-validation for the one-stage study.

		Predicted classes Normal	Benign	Malignant
Actual classes	Normal	862	91	59
	Benign	123	889	0
	Malignant	166	2	844

Table 5: The evaluation metrics of each classifier evaluated by elevenfold cross-validation for the one-stage study.

Classifier	SNS	SPC	PPV	NPV	FPR	FNR	FDR	FOR	ACC
LLC	85.47	92.74	85.47	92.74	7.26	14.53	14.53	7.26	85.47
FLDA	81.72	90.86	81.72	90.86	9.14	18.28	18.28	9.14	81.72
LDC	80.80	90.40	80.80	90.40	9.60	19.20	19.20	9.60	80.80
SVM (‘p’, 1)	77.14	88.57	77.14	88.57	11.43	22.86	22.86	11.43	77.14
Decision tree	77.50	88.75	77.50	88.75	11.25	22.50	22.50	11.25	77.50
Random forest	75.86	87.93	75.86	87.93	12.07	24.14	24.14	12.07	75.86
Naïve Bayes	65.42	82.71	65.42	82.71	17.29	34.58	34.58	17.29	65.42
kNN ($k = 5$)	42.13	71.06	42.13	71.06	28.94	57.87	57.87	28.94	42.13

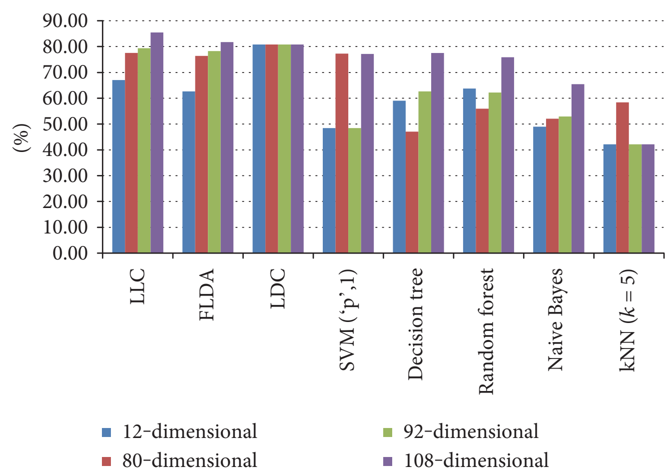


Figure 4: Average classification accuracies of eight different classifiers obtained by elevenfold cross-validation for the one-stage study using different feature sets.

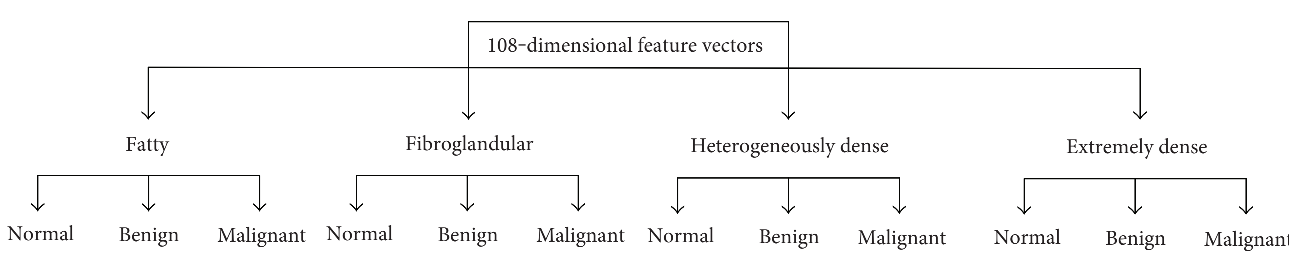


Figure 5: Flowchart designed for the two-stage study.

cross-validation for the two-stage study are given in Tables 6 and 7, respectively. Similar results are obtained in the two-stage study as in the one-stage study. The confusion matrices in Tables 6 and 7 clearly show that the false negatives and false positives for both benign and malignant classes belong to the normal class. The terms N., B., and M. in Table 6 refer to the normal, benign and malignant classes, respectively.

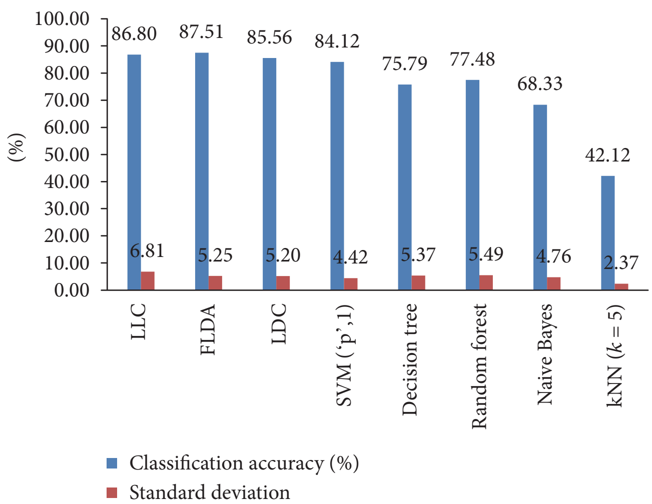


Figure 6: Average classification accuracies and standard deviations of classifiers obtained by elevenfold cross-validation for the two-stage study.

The evaluation metrics of each classifier and the classifier combination evaluated by elevenfold cross-validation for the two-stage study are given in Tables 8 and 9, respectively.

3.1.3. Three-Stage Study. After the classification accuracies are enhanced by the two-stage study, the authors propose a three-stage study for further improvement. The three-stage study consists of both breast tissue and health status classification, where the health status classification is achieved through two consecutive stages. In the first stage of this study, the feature vectors are classified into breast tissue classes similar to those in the two-stage study. The breast-tissue-type-defined feature vectors are then categorized into normal and abnormal classes in the second stage. Finally, in the last stage, the feature vectors labeled as abnormal classes are categorized into benign and malignant classes. The flowchart for the three-stage study is illustrated in Figure 7.

The average classification accuracies and standard deviations of eight classifiers obtained by elevenfold cross-validation technique for the three-stage study are graphically shown in Figure 8. The FLDA has the best classification performance with a maximum of 93.29% accuracy rate among all classifiers. In this case, as the prior probabilities of the classes are not equal again as in the two-stage study, the classification success of the LLC classifier is less than that of the FLDA and LDC classifiers.

The total confusion matrices of the (a) FLDA, (b) LDC, and (c) LLC classifiers obtained by elevenfold cross-validation for the three-stage study, and the total confusion matrix of classifier combination obtained by elevenfold cross-validation for the three-stage study are given in Tables 10 and 11, respectively. In the three-stage study, as seen in the tables, mammograms in normal and benign classes are exactly inseparable from each other, while malignant mammograms are clearly distinguished from the normal and benign classes. The terms N., B., and M. in Table 10 stand for the normal, benign, and malignant classes, respectively.

Table 6: Total confusion matrices of the (a) FLDA, (b) LLC, and (c) LDC classifiers obtained by elevenfold cross-validation for the two-stage study.

(a)

		Predicted classes		
		N.	B.	M.
Actual classes	N.	872	52	88
	B.	186	826	0
	M.	142	5	865

(b)

		Predicted classes		
		N.	B.	M.
Actual classes	N.	835	105	72
	B.	135	874	3
	M.	153	0	859

(c)

		Predicted classes		
		N.	B.	M.
Actual classes	N.	877	25	108
	B.	218	794	0
	M.	228	2	782

Table 7: Total confusion matrix of classifier combination obtained by elevenfold cross-validation for the two-stage study.

		Predicted classes		
		Normal	Benign	Malignant
Actual classes	Normal	887	41	84
	Benign	162	850	0
	Malignant	140	2	870

If Figure 8 is carefully examined, the FLDA, LDC, and LLC classifiers, as in the two-stage study, are the best three classifiers in terms of recognition accuracy. The results of these classifiers are combined via majority voting and eventually the classification performance is increased to 93.52%.

The evaluation metrics of each classifier and the classifier combination evaluated by elevenfold cross-validation for the three-stage study are given in Tables 12 and 13, respectively.

3.2. Discussion. The proposed feature ensemble is formed by concatenating the LCP-based, statistical, and frequency-domain features. The LCP algorithm is performed by itself for several image processing applications. The motivation behind the usage of the LCP algorithm for feature extraction relies on the decomposition of information existing in breast mammogram images. Moreover, the LCP features include pixel-wise relationships. As it covers relatively few relationships among pixels in a breast mammogram image, the LCP is used as the fundamental feature extraction method

Table 8: The evaluation metrics of each classifier evaluated by elevenfold cross-validation for the two-stage study.

Classifier	SNS	SPC	PPV	NPV	FPR	FNR	FDR	FOR	ACC
LLC	86.49	92.11	84.92	91.78	7.89	13.51	15.08	8.22	86.80
FLDA	87.45	92.18	85.25	91.76	7.82	12.55	14.75	8.24	87.51
LDC	84.68	90.29	81.66	89.41	9.71	15.32	18.34	10.59	85.56
SVM (‘p’, 1)	84.46	90.13	80.30	89.51	9.87	15.54	19.70	10.49	84.12
Decision tree	75.78	87.64	74.93	88.00	12.36	24.22	25.07	12.00	75.79
Random forest	77.51	88.29	77.38	88.38	11.71	22.49	22.62	11.62	77.48
Naïve Bayes	67.81	83.86	71.38	84.75	16.14	32.19	28.62	15.25	68.33
kNN ($k = 5$)	42.13	70.51	40.73	70.81	29.49	57.87	59.27	29.19	42.12

Table 9: The evaluation metric of classifier combination evaluated by elevenfold cross-validation for the two-stage study.

Classifier	SNS	SPC	PPV	NPV	FPR	FNR	FDR	FOR	ACC
Classifier combination	88.67	92.93	86.32	92.45	7.07	11.33	13.68	7.55	88.79

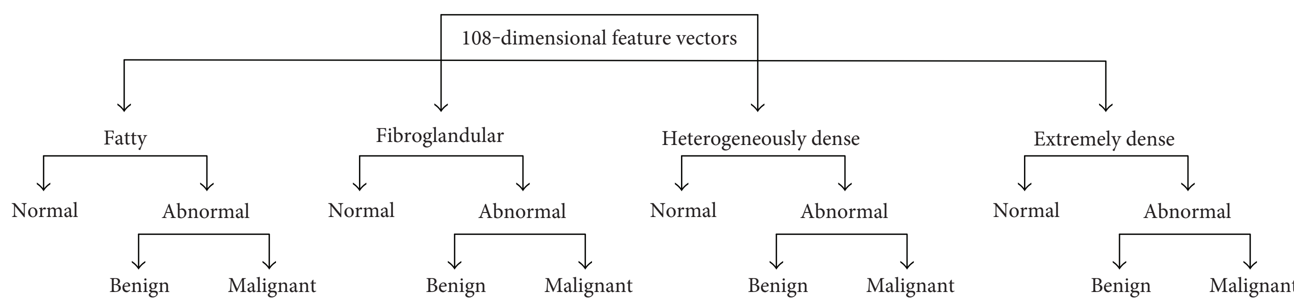


Figure 7: Flowchart designed for the three-stage study.

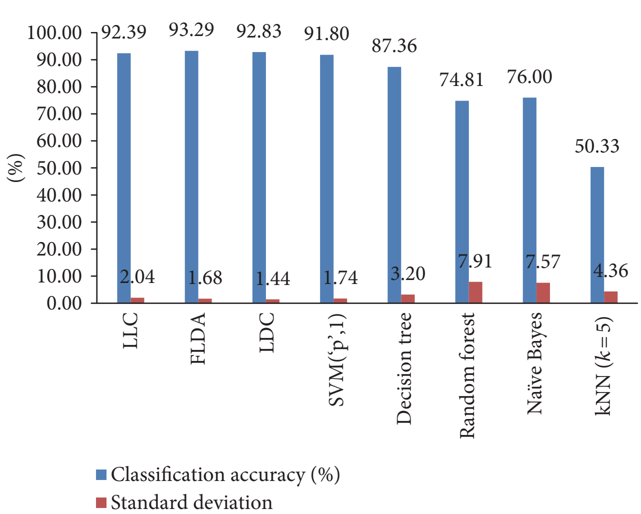


Figure 8: Average classification accuracies and standard deviations of classifiers obtained by elevenfold cross-validation for the three-stage study.

to explore the underlying information in an image. However, the LCP features are not completely adequate to efficiently classify mammogram parts because it can be affected by various issues. Therefore, the use of LCP only will not result in the most representative features for a mammogram. Furthermore, twelve statistical features were calculated from the LCP features. The positive impact of statistical features extracted directly from the image texture on classification success is already known [52]. In addition, the LCP feature vectors extracted from breast mammograms are indicated as successfully discriminative features [5]. Hence, in this study, the statistical features are obtained from the LCP feature vectors rather than directly from the mammogram image pixel matrices. Moreover, 16 frequency-domain features are computed and appended to other two types of features (LCP-based and statistical features). Since the brightness is one of the most significant issues for breast cancer diagnosis and the variations of brightness in a mammogram image can be obviously observed in spatial domain, it is assumed that frequency-domain features are also representative of mammograms in this study. Ultimately, the feature vectors that have more representative power and are more robust to numerous effects are constructed by this method.

Additionally, a multistage classification scheme is proposed in this study. It consists of three cases: the one-stage study, two-stage study, and three-stage study. In the one-stage study, the feature vectors are classified according to only their health status regardless of the breast tissue type

Table 10: Total confusion matrices of the (a) FLDA, (b) LDC, and (c) LLC classifiers obtained by elevenfold cross-validation for the three-stage study.

(a)

		Predicted classes		
		N.	B.	M.
Actual classes	N.	393	445	174
	B.	174	629	209
	M.	139	260	613

(b)

		Predicted classes		
		N.	B.	M.
Actual classes	N.	382	490	140
	B.	183	697	132
	M.	204	311	497

(c)

		Predicted classes		
		N.	B.	M.
Actual classes	N.	336	467	209
	B.	145	624	243
	M.	158	249	605

of mammograms. The standard deviation values for the one-stage study are high since some folds in cross-validation process provide high recognition accuracies but the other folds give much lower classification accuracies. This situation clearly implies that the accuracy results of the one-stage study are directly related with the mammogram parts used in train/test separation of each fold. If a test set includes more similar parts compared to those in the corresponding train set, the accuracy suddenly raises. On the contrary, if the similarities between the test and train sets are weak, the classification fails. This consequence obviously reveals that the one-stage study does not give trustworthy accuracy results. In order to prevent the high standard deviation problem and increase the classification accuracy rates, the two-stage study is implemented. In the two-stage study, both breast tissue and health status classification are consecutively performed. By this way, the breast tissue types of mammograms are taken into consideration so that a more reliable classification is achieved. The trustworthiness of recognition can be inferred by examining the standard deviation values for each classifier. These values are much lower compared to those obtained in the one-stage study. Therefore, the accuracy results of the two-stage study are not related with the mammogram parts treated in train/test separation of each fold. The cross-validation process gives more reliable accuracy rates. Finally, the three-stage study considers both breast tissue and health status classification as the two-stage study does, except that the health status classification is realized through two consequent stages. By this way, the lowest standard deviation values especially for the classifiers which give higher recognition accuracies are obtained. This outcome apparently

Table 11: Total confusion matrix of classifier combination obtained by elevenfold cross-validation for the three-stage study.

		Predicted classes		
		Normal	Benign	Malignant
Actual classes	Normal	393	444	175
	Benign	174	630	208
	Malignant	139	257	616

exposes that the three-stage study not only performs the most reliable classification process but also is independent from mammogram parts used in training and test sets of each cross-validation fold. Besides, the most successful experiments are achieved in the three-stage case. Ultimately, if one considers both success and reliability issues at the same time in this classification problem, the three-stage case provide these two issues simultaneously.

The mammogram parts of fatty breast tissue type in the IRMA database are classified using only LCP-based feature vectors, and a maximum of 90.60% recognition accuracy is attained in [5]. By the proposed feature ensemble and multistage classification, this accuracy is effectively increased to 93.52% for all tissue types rather than for only one breast tissue type. This result explicitly shows that the new feature ensemble is more representative than an LCP-based feature vector by itself, and the proposed multistage classification scheme is more successful and reliable than a single-stage classification for breast cancer diagnosis. The comparison of the proposed study with other studies in the literature is given in Table 14.

4. Conclusion

Breast cancer is the second major reason for female deaths resulting from cancer worldwide. Although there is no known way to prevent breast cancer, mortality can be reduced only with early diagnosis. Therefore, the computer-aided diagnosis (CAD) systems are very important as they allow radiologists to reconsider mammogram images with increased sensitivity of detection and diagnosis. In this study, a multistage classification scheme using a novel and discriminative feature ensemble to be implemented in a CAD system for breast cancer diagnosis is proposed. The proposed system is verified using the IRMA database. This database includes all twelve classes defined by BI-RADS, which are four different breast tissue types, and three different health status cases for each breast tissue type. The proposed feature ensemble is formed by concatenating the 80-dimensional LCP-based features obtained from the one-level, two-dimensional discrete wavelet transform of the preprocessed mammogram images, 12-dimensional statistical features computed from the LCP-based features, and 16-dimensional frequency-domain features calculated from the two-level two-dimensional discrete wavelet transform of the preprocessed mammogram images. In this study, a multistage classification scheme, namely the one-stage study, two-stage study, and three-stage study cases, is presented. The feature vectors are classified directly according to their health status in the one-stage study. In the two-stage study, the health status classification of each

Table 12: The evaluation metrics of each classifier evaluated by elevenfold cross-validation for the three-stage study.

Classifier	SNS	SPC	PPV	NPV	FPR	FNR	FDR	FOR	ACC
LLC	87.90	93.25	90.73	94.05	6.75	12.10	9.27	5.95	92.39
FLDA	87.91	92.90	90.61	93.96	7.10	12.09	9.39	6.04	93.29
LDC	87.30	91.71	88.29	92.17	8.29	12.70	11.71	7.83	92.83
SVM (‘p’, 1)	74.85	85.29	79.66	88.45	14.71	25.15	15.43	11.55	91.80
Decision tree	85.68	91.60	87.03	92.18	8.40	14.32	12.97	7.82	87.36
Random forest	79.54	88.69	78.44	89.24	11.31	20.46	21.56	10.76	74.81
Naïve Bayes	65.23	82.04	68.49	85.73	17.96	34.77	31.51	14.27	76.00
kNN ($k = 5$)	63.02	77.98	54.81	78.50	22.02	36.98	45.19	21.50	50.33

Table 13: The evaluation metric of classifier combination evaluated by elevenfold cross-validation for the three-stage study.

Classifier	SNS	SPC	PPV	NPV	FPR	FNR	FDR	FOR	ACC
Classifier combination	87.91	92.90	90.61	93.96	7.10	12.09	9.39	6.04	93.52

Table 14: The comparison of the results for the proposed CAD system.

Authors	Features	Number of images	Accuracy
Ganesan et al. [21]	Statistical features	300	91%
Korkmaz and Korkmaz [65]	Statistical features	378	98.3%
Vikhe and Thool [66]	Unstated	130	91%
Jen and Yu [67]	Statistical and gradient features	322	86%
Vadivel and Surendiran [68]	Shape and margin	224	87.76%
Acharya et al. [69]	Area, homogeneity	360	88.80%

breast tissue type, determined in the first stage where the breast tissue classification is achieved, is executed. The three-stage study also considers both breast tissue and health status; however, in this case, the health status classification is performed with two consequent stages, where the normal and abnormal mammograms are determined first, and the abnormal defined mammograms are then classified as benign and malignant. The maximum recognition accuracy of the proposed system is obtained in the three-stage study. These results clearly indicate that using three-stage study is very effective for a CAD system and helpful for radiologists to make more accurate breast cancer diagnoses.

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 Scientific Research Project Coordination Unit of Eskişehir Osmangazi University (Project no.: 201515D10). The database utilized in this study was used by the courtesy of Thomas M. Deserno, Department of Medical Informatics, Division of Image and Data Management, Aachen, Germany.

References

[1] American Cancer Society, *Cancer Facts & Figures 2015*, American Cancer Society, Atlanta, 2015.

[2] American Cancer Society, *Cancer Facts & Figures 2009*, American Cancer Society, Atlanta, 2009.

[3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics,” *CA: a Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[4] American Cancer Society, *Breast Cancer Facts & Figures 2013-2014*, American Cancer Society, Inc., Atlanta, 2013.

[5] S. Ergin and O. Kılınç, “A new feature extraction framework based on wavelets for breast cancer diagnosis,” *Computers in Biology and Medicine*, vol. 51, pp. 171–182, 2014.

[6] R. L. Birdwell, D. M. Ikeda, K. F. O’Shaughnessy, and E. A. Sickles, “Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection,” *Radiology*, vol. 219, no. 1, pp. 192–202, 2001.

[7] H. Al-Ghaib, R. Adhami, and M. Scott, “An overview of mammogram analysis,” *IEEE Potentials*, vol. 35, pp. 21–28, 2016.

[8] N. Al-Najdawi, M. Biltawi, and S. Tedmori, “Mammogram image visual enhancement, mass segmentation and classification,” *Applied Soft Computing*, vol. 35, no. 6, pp. 175–185, 2015.

[9] Y. M. George, H. H. Zayed, M. I. Roushdy, and B. M. Elbagoury, “Remote computer-aided breast cancer detection and diagnosis system based on cytological images,” *IEEE Systems Journal*, vol. 8, no. 3, pp. 949–964, 2014.

[10] S. C. Tai, Z. S. Chen, and W. T. Tsai, “An automatic mass detection system in mammograms based on complex texture features,” *Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 618–627, 2014.

[11] Z. Yang, M. Dong, Y. Guoa et al., "A new method of microcalcifications detection in digitized mammograms based on improved simplified PCNN," *Neurocomputing*, vol. 218, pp. 79–90, 2016.

[12] J. Anitha and J. Dinesh Peter, "Mammogram segmentation using maximal cell strength updation in cellular automata," *Medical & Biological Engineering & Computing*, vol. 53, no. 8, pp. 737–749, 2015.

[13] M. Husaain, "Mammogram enhancement using lifting dyadic wavelet transform and normalized Tsallis entropy," *Journal of Computer Science and Technology*, vol. 29, no. 6, pp. 1048–1057, 2014.

[14] J. N. Silva, A. O. C. Filho, A. C. Silva, A. C. Paiva, and M. Gattass, "Automatic detection of masses in mammograms using quality threshold clustering, correlogram function, and SVM," *Journal of Digital Imaging*, vol. 28, no. 3, pp. 323–337, 2015.

[15] B. Gupta and M. Tiwari, "A tool supported approach for brightness preserving contrast enhancement and mass segmentation of mammogram images using histogram modified grey relational analysis," *Multidimensional Systems and Signal Processing*, pp. 1–19, 2016.

[16] M. J. Lado, C. Cadarso-Suárez, J. Roca-Pardiñas, and P. G. Tahoces, "Categorical variables, interactions and generalized additive models. Applications in computer-aided diagnosis systems," *Computers in Biology and Medicine*, vol. 38, no. 4, pp. 475–483, 2008.

[17] H. B. Kekre, T. Sarode, and S. Gharge, "Tumor detection in mammography images using vector quantization technique," *International Journal of Intelligent Information Technology Application*, vol. 2, pp. 237–242, 2009.

[18] H. B. Kekre, T. Sarode, S. Gharge, and K. Raut, "Detection of cancer using vector quantization for segmentation," *International Journal of Computers and Applications*, vol. 4, pp. 14–19, 2010.

[19] W. Haider, M. Sharif, and M. Raza, "Achieving accuracy in early stage tumor identification systems based on image segmentation and 3d structure analysis," *Computer Engineering and Intelligent Systems*, vol. 2, pp. 96–103, 2011.

[20] M. Radovic, M. Djokovic, A. Peulic, and N. Filipovic, "Application of data mining algorithms for mammogram classification," in *Proceedings of the 13th IEEE International Conference on Bioinformatics and Bioengineering*, pp. 1–4, 2013.

[21] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, B. Matthew, and A. K. Thomas, "Decision support system for breast cancer detection using mammograms," *Proceedings of the Institution of Mechanical Engineers. Part H*, vol. 227, no. 7, pp. 721–732, 2013.

[22] J. B. Li, Y. H. Wang, S. C. Chu, and J. F. Roddick, "Kernel self-optimization learning for kernel-based feature extraction and recognition," *Information Sciences*, vol. 257, pp. 70–80, 2014.

[23] R. P. Ramos, M. Z. do Nascimento, and D. C. Pereira, "Texture extraction: an evaluation of ridgelet, wavelet and co-occurrence based techniques applied to mammograms," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11036–11047, 2012.

[24] B. Shradhananda, M. Banshidhar, and D. Ratnakar, "Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer," *Neurocomputing*, vol. 154, no. Issue C, pp. 1–4, 2015.

[25] N. Vallez, G. Bueno, O. Deniz et al., "Breast density classification to reduce false positives in CADe systems," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 569–584, 2013.

[26] S. K. Biswas and D. P. Mukherjee, "Recognizing architectural distortion in mammogram: a multiscale texture modeling approach with GMM," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 7, pp. 2023–2030, 2011.

[27] N. R. Pal, B. Bhowmick, S. K. Patel, S. Pal, and J. Das, "A multistage neural network aided system for detection of microcalcifications in digitized mammograms," *Neurocomputing*, vol. 71, no. 13, pp. 2625–2634, 2008.

[28] Z. Chen, H. Strange, A. Oliver, E. R. E. Denton, C. Boggis, and R. Zwiggelaar, "Topological modeling and classification of mammographic microcalcification clusters," *IEEE Tractions on Bio-Medical Engineering*, vol. 62, no. 4, pp. 1203–1214, 2015.

[29] A. Papadopoulos, D. I. Fotiadis, and L. Costaridou, "Improvement of microcalcification cluster detection in mammography utilizing image enhancement techniques," *Computers in Biology and Medicine*, vol. 38, no. 10, pp. 1045–1055, 2008.

[30] P. Agrawal, M. Vatsa, and R. Singh, "Saliency based mass detection from screening mammograms," *Signal Processing*, vol. 99, pp. 29–47, 2014.

[31] F. Moayedi, Z. Azimifar, R. Boostani, and S. Katebi, "Contourlet-based mammography mass classification using the SVM family," *Computers in Biology and Medicine*, vol. 40, no. 4, pp. 373–383, 2010.

[32] A. N. Karahaliou, I. S. Boniatis, S. G. Skiadopoulos et al., "Breast cancer diagnosis: analyzing texture of tissue surrounding microcalcifications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 6, pp. 731–738, 2008.

[33] W. B. Sampaio, E. M. Diniz, A. C. Silva, A. C. de Paiva, and M. Gattass, "Detection of masses in mammogram images using CNN, geostatistic functions and SVM," *Computers in Biology and Medicine*, vol. 41, no. 8, pp. 653–664, 2011.

[34] A. Keleş, A. Keleş, and U. Yavuz, "Expert system based on neuro-fuzzy rules for diagnosis breast cancer," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5719–5726, 2011.

[35] M. M. R. Krishnan, S. Banerjee, C. Chakraborty, C. Chakraborty, and A. K. Ray, "Statistical analysis of mammographic features and its classification using support vector machine," *Expert Systems with Applications*, vol. 37, no. 1, pp. 470–478, 2010.

[36] B. Verma, P. McLeod, and A. Klevansky, "A novel soft cluster neural network for the classification of suspicious areas in digital mammograms," *Pattern Recognition*, vol. 42, no. 9, pp. 1845–1852, 2009.

[37] A. Oliver, A. Torrent, X. Liado et al., "Automatic microcalcification and cluster detection for digital and digitized mammograms," *Knowledge-Based Systems*, vol. 28, pp. 68–75, 2012.

[38] X. Zhang and X. Gao, "Twin support vector machines and subspace learning techniques for microcalcification clusters detection," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 5, pp. 1062–1072, 2012.

[39] K. Ganesan, U. R. Acharya, C. K. Chua, C. M. Lim, and K. T. Abraham, "One-class classification of mammograms using trace transform functions," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 2, pp. 304–311, 2014.

[40] E. Malar, A. Kandaswamy, D. Chakravarthy, and A. GiriDharan, "A novel approach for detection and classification of mammographic microcalcifications using wavelet analysis and extreme learning machine," *Computers in Biology and Medicine*, vol. 42, no. 9, pp. 898–905, 2012.

[41] A. Bria, N. Karssemeijer, and F. Tortorella, "Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications," *Medical Image Analysis*, vol. 18, pp. 241–252, 2014.

[42] M. Hachama, A. Desolneux, and F. J. P. Richard, "Bayesian technique for image classifying registration," *IEEE Transactions on image processing*, vol. 21, no. 9, pp. 4080–4091, 2012.

[43] R. Savitha, S. Suresh, and N. Sundararajan, "Projection-based fast learning fully complex-valued relaxation neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 529–541, 2013.

[44] J. A. Ventura and J. M. Chen, "Segmentation of two-dimensional curve contours," *Pattern Recognition*, vol. 25, no. 10, pp. 1129–1140, 1992.

[45] D. Guliato, R. M. Rangayyan, J. D. Carvalho, and S. A. Santigo, "Polygonal modeling of contours of breast tumors with the preservation of spicules," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 14–20, 2008.

[46] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[47] Y. Guo, G. Zhao, and M. Pietkäinen, "Texture classification using a linear configuration model based descriptor," *British Machine Vision Association*, vol. 119, pp. 1–10, 2011.

[48] İ. Işıklı Esener, S. Ergin, and T. Yüksel, "A new ensemble of features for breast cancer diagnosis," in *Proc 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1168–1173, May 2015.

[49] T. M. Deserno, J. E. E. de Oliveira, and A. A. Araujo, "Towards computer-aided diagnostics of screening mammography using content-based image retrieval," in *24th SIBGRAPI Conference on Graphics, Patterns and Images (Sibgrapi)*, pp. 211–219, August 2011.

[50] T. Ahonen, A. Hadid, and M. Pietkäinen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[51] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor," *IEEE Biometrics Compendium*, vol. 19, no. 2, pp. 533–544, 2010.

[52] G. Zhang, X. Huang, S. Z. Li, Y. Wang, and X. Wu, "Boosting local binary pattern (LBP)-based face recognition," *Lecture Notes in Computer Science*, vol. 3338, pp. 179–186, 2005.

[53] T. Ojala, M. Pietkäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[54] K. Woods, C. Doss, K. Bowyer, J. Solka, C. Priebe, and W. Philip, "Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 6, pp. 1417–1435, 1993.

[55] I. Daubechies, "Ten lectures on wavelets," in *CBMS-NSF Regional Conference Series in Applied Mathematics*, p. 194, 1992.

[56] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[57] A. R. Webb, "Linear discriminant analysis," in *Statistical Pattern Recognition*, pp. 123–124, John Wiley & Sons, New York, 2002.

[58] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055–1064, 1999.

[59] K. Özkan, S. Ergin, Ş. Işık, and İ. Işıklı, "A new classification scheme of plastic wastes based upon recycling labels," *Waste Management*, vol. 35, pp. 29–35, 2015.

[60] A. R. Webb, "Linear discriminant analysis," in *Statistical Pattern Recognition*, pp. 158–159, John Wiley & Sons, New York, 2002.

[61] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[62] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[63] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with naive Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432–5435, 2009.

[64] E. Fix and J. Hodges, *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties, International Statistical Review (Revue Internationale de Statistique)*, vol. 57, pp. 238–247, 1989.

[65] S. A. Korkmaz and M. F. Korkmaz, "A new method based cancer detection in mammogram textures by finding feature weights and using Kullback-Leibler measure with kernel estimation," *Optik*, vol. 126, no. 20, pp. 2576–2583, 2015.

[66] P. S. Vikhe and V. R. Thool, "Mass detection in mammographic images using wavelet processing and adaptive threshold technique," *Journal of Medical Systems*, vol. 40, no. 4, pp. 82–97, 2016.

[67] C. C. Jen and S. S. Yu, "Automatic detection of abnormal mammograms in mammographic images," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3048–3055, 2015.

[68] A. Vadivel and B. Surendiran, "A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories," *Computers in Biology and Medicine*, vol. 43, no. 4, pp. 259–267, 2013.

[69] U. R. Acharya, E. Y. K. Ng, Y. Hong, Y. Jie, and G. J. L. Kaw, "Computer-based identification of breast cancer using digitized mammograms," *Journal of Medical Systems*, vol. 32, no. 6, pp. 499–507, 2008.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 1439643, 8 pages
https://doi.org/10.1155/2017/1439643

Research Article

New Directions in 3D Medical Modeling: 3D-Printing Anatomy and Functions in Neurosurgical Planning

Paolo Gargiulo,[1] Íris Árnadóttir,[1] Magnús Gíslason,[1] Kyle Edmunds,[1] and Ingvar Ólafsson[2]

[1]*Institute of Biomedical and Neural Engineering/Medical Technology Center, Reykjavik University and Landspitali University Hospital, Menntavegi 1, 101 Reykjavik, Iceland*
[2]*Department of Neurosurgery, Landspitali University Hospital, Áland, 108 Reykjavik, Iceland*

Correspondence should be addressed to Paolo Gargiulo; paologar@landspitali.is

Received 2 March 2017; Accepted 13 April 2017; Published 8 June 2017

Academic Editor: Pan Lin

This paper illustrates the feasibility and utility of combining cranial anatomy and brain function on the same 3D-printed model, as evidenced by a neurosurgical planning case study of a 29-year-old female patient with a low-grade frontal-lobe glioma. We herein report the rapid prototyping methodology utilized in conjunction with surgical navigation to prepare and plan a complex neurosurgery. The method introduced here combines CT and MRI images with DTI tractography, while using various image segmentation protocols to 3D model the skull base, tumor, and five eloquent fiber tracts. This 3D model is rapid-prototyped and coregistered with patient images and a reported surgical navigation system, establishing a clear link between the printed model and surgical navigation. This methodology highlights the potential for advanced neurosurgical preparation, which can begin before the patient enters the operation theatre. Moreover, the work presented here demonstrates the workflow developed at the National University Hospital of Iceland, Landspitali, focusing on the processes of anatomy segmentation, fiber tract extrapolation, MRI/CT registration, and 3D printing. Furthermore, we present a qualitative and quantitative assessment for fiber tract generation in a case study where these processes are applied in the preparation of brain tumor resection surgery.

1. Introduction

Three-dimensional (3D) modeling and rapid prototyping technologies have recently shown great utility in a wide variety of applications in medicine and surgery [1, 2]. In principle, the 3D recapitulation of patient-specific anatomical features provides surgeons with an immediate and intuitive understanding of even the most complex anatomical morphologies, enabling accurate planning and emulation of a host of surgical procedures [3, 4]. Indeed, the employment of these 3D anatomical models is additionally being considered for a host of implantation procedures, such as dental crowning, craniofacial reconstruction, and tissue regeneration via biological scaffolds [5–7].

Kodama et al. reported the birth of 3D rapid prototyping in 1981 [8], and the first use of the technology in support of surgical planning was reported by Anderl et al. in 1994 [9]. Since then, improvements in medical imaging modalities, such as CT and MRI, have driven both the clinical interest and academic development of 3D rapid prototyping in a medical context. Modern rapid prototyping enables the construction of anatomical models with layer thicknesses on the order of microns, and with concurrent advancements in medical image contrast segmentation, these models are able to recapitulate external and internal anatomical morphologies to high degrees of precision. The utilization of rapid prototyping models incurs a host of benefits to many surgical fields, which include improving surgical planning, enhancing diagnostic quality, decreasing patient exposure time to general anesthesia, decreasing patient blood loss, and shortening wound exposure time [10].

With the aims of improving surgical outcomes, reducing future costs, and developing thorough clinical guidelines for enhancing surgical planning and assessment, the National University Hospital of Iceland, Landspitali, established an in-house service for 3D rapid prototyping in 2007. Since its introduction, this service has allowed physicians and surgeons from different specialities to submit requests for a host

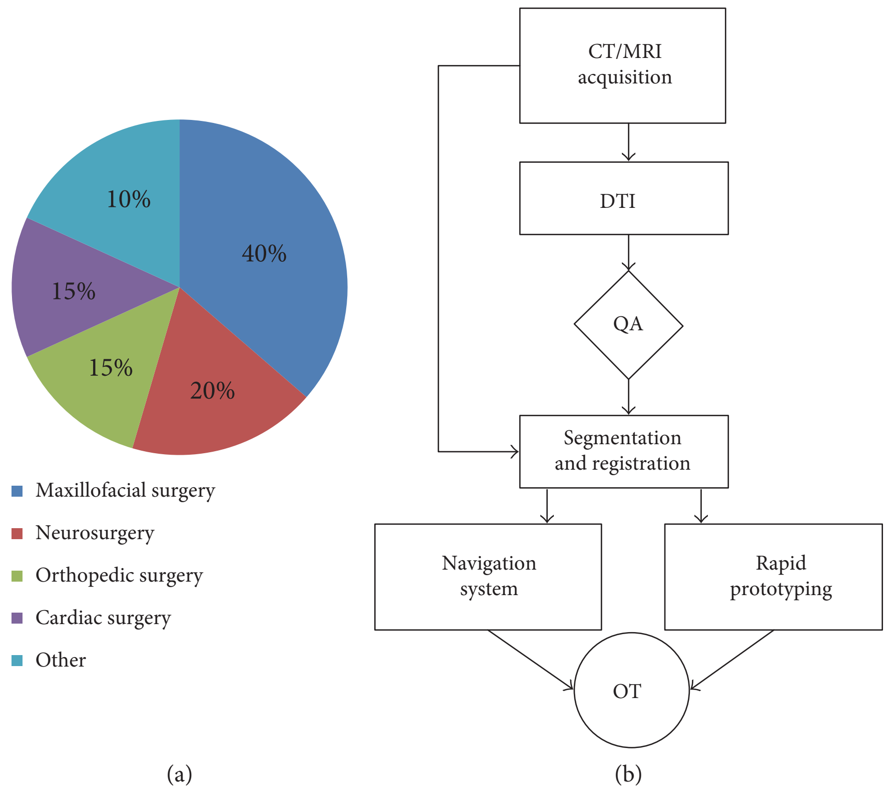


Figure 1: Clinical areas associated with the 200 surgeries assisted with 3D-printed models (a). Block diagrams showing the different steps required to create a 3D-printed model based on CT, MRI, and DTI data.

of 3D models to be made available within 24 hours of submission. This process was simultaneously employed in research activities to study both the anthropometry of human muscles [11] and the use of rapid prototyping as preparation for complex brain surgeries in combination with neurosurgical navigation systems [12]. Since then, the National University Hospital of Iceland has fabricated over 200 surgical models for patient cases in the fields of cardiac, orthopedic, and neurosurgery (Figure 1(a)). The overwhelming success of this 3D rapid prototyping service has led to its solidification as an essential service within the hospital, and the rapid prototyping service continues to expand its impact on an increasing number of assisted surgical cases [13].

In neurosurgery, one technique that has likewise been increasingly used for preoperative planning is diffusion tensor imaging (DTI) tractography or fiber tracking. Tractography is a noninvasive technique that allows for the in vivo localization of fiber tracts in the brain. Tractography uses DTI, which is based on magnetic resonance imaging (MRI), to map brain connectivity, which can provide neurosurgeons with the opportunity to visualize nerve fiber tracts before surgery [14]. More specifically, this technique may be applied to that of functional MRI (fMRI), which has shown great utility in the context of surgical planning. fMRI utilizes hemodynamic responses within the brain to implicate regional recruitment with a variety of cortical functions, such as motor control and language processing [15]. Unfortunately, the routine integration of surgical planning, DTI, and preoperative fMRI has been primarily limited by concerns regarding acquisition and registration reliability. Nonetheless, there is much promise in this regard—evidenced most relevantly in the reconstruction of cotricospinal tracts for preoperative tumor planning [16, 17]. The purpose of this paper is to detail a novel approach to neurosurgical planning via the use of 3D printing, which combines patient-specific anatomy from traditional computer tomography (CT) and MRI images, with brain function derived from the in vivo localization of fiber tracts in the brain using DTI.

2. Material and Methods

The procedure of creating the 3D-printed models based on CT, MRI, and DTI data can be seen in Figure 1(b).

2.1. CT and MRI Acquisition. CT data were acquired from a Philips/Brilliance 64, the Head scan protocol was set to 119 mA X-ray for the tube current and 120 KV for tube voltage, and the slice thickness is typically between 0.6 and 1 mm.

MRI data were acquired from a 1.5T Siemens Avanto and the head coil used was Head Matrix Coil from Siemens. Both anatomical images and DTI were acquired for this process. The DTI protocol included a spin echo-echo planar imaging- (SE-EPI-) based DTI sequence with 20 diffusion directions, two repetitions to boost the SNR and b value (b is the diffusion sensitivity) equal to 1000 s/mm^2. The anatomical image protocol included a T1-weighted 3D magnetization-prepared rapid acquisition gradient echo sequence (MP-RAGE).

Table 1: Grades for incorrectly displayed fibers.

Incorrectly displayed fibers	
Grade 1	None
Grade 2	<10% of all displayed fibers
Grade 3	<25% of all displayed fibers
Grade 4	>50% of all displayed fibers

Table 2: Grades for anatomic accuracy.

Anatomic accuracy	
Grade 1	Follow fiber tracts within anatomical boundaries
Grade 2	Follow fiber tracts outside anatomical boundaries
Grade 3	Follow poorly anatomical fiber tracts
Grade 4	Do not follow anatomical fiber tracts

2.2. DTI: Fiber Tract Extrapolation. Two different software (StealthViz [18] and nordicBrainEx [19]) programs were used to extrapolate the optimal fiber tracts for planning and rapid prototyping.

The fiber tracts of major interest for this process are the so called *eloquent fiber tracts*; these tracks are easily clinically assessed and are most important for the patient outcome. In total, five fiber tracts were extrapolated from both software platforms:

(i) Corpus callosum: the corpus callosum is located in the center of the brain and forms the largest white matter bundle. Its role is to transfer information between the left and right cerebral hemispheres [20].

(ii) Motor tracts: they originate in the motor cortex area and descends down to the brain stem and spinal cord to control α-motor neurons. They can control posture, reflexes, and muscle tone as well as conscious voluntary movements [21].

(iii) Sensory tracts: they are responsible for the sense of touch. They receive incoming messages for touch and limb movements from the body [22].

(iv) Optic tracts: they transfer the information from the retina to the visual cortex of the brain [23, 24].

(v) Broca's area to Wernicke's area: arcuate Fasciculus is the prominent fiber tracts that connect these two areas that play a role in our language and speech [25].

2.2.1. StealthViz. StealthViz is a surgical planning software application. It allows import of Digital Imaging and Communications in Medicine (DICOM) datasets that can be reviewed in 2D and with 3D volume rendering, multimodality image fusion, and segmentation of structures with manual and semiautomatic tools. The software performs white matter tractography. It enables realignment of diffusion-weighted gradient, coregistration with other anatomical and functional datasets, and tensor calculations. The fiber tracking uses deterministic FACT algorithm [26]. The workflow is the following:

(i) Import data: the MRI data are imported in DICOM format. Anatomical and diffusion tensor images are merged and the diffusion tensor positioned in the correct anatomical position.

(ii) Segmentation: StealthViz allows segmentation with five different tools; pick region tool, brush tool, lasso tool, magic wand, and blow. A brain tumor can be segmented by using a *blow tool* which marks the region of interest on one cross section. The process can be iterated on several slices and those marked regions can be interpolated creating a 3D object of the tumor.

(iii) Fiber tracking: to trace tracts in StealthViz a *start box* (and eventually a middle and end box) can placed on specific regions of interest in the brain, called seeding point, for example, in our application, we start in the region of corpus callosum. Then, the software computed all the fibers that go through the designed *box*. Different combinations of the boxes can be used to find the tracts of interest. Tracks that are not of interest can be removed. Calculated fiber tracts are visualized within the structural images both in 2D and in 3D.

(iv) Calculate as 3D object: when the tractography planning is completed and approved by neurosurgeon, the tracts are converted in 3D objects and saved in a DICOM format. In this phase, an error margin of 1 mm is added to each fiber tract.

(v) Export planning: results can be exported as a one file or separated files (each for every track) to the surgical navigation system or to a USB flash memory.

2.2.2. NordicBrainEx. NordicBrainEx is DICOM compatible and can analyze DTI data acquired with all major MRI scanners. DTI datasets acquired with two different b values (one $b = 0$ and six or more DWI where $b \neq 0$) can be analyzed in nordicBrainEx. The DTI analysis in nordicBrainEx generates parametric maps of various attributes of the diffusion tensor, including eigenvector color map (cDTI), fractional anisotropy index (FA), mean diffusivity (ADC), tensor eigenvalues ($\lambda_1, \lambda_2,$ and λ_3), and trace weighted (TraceW). The fiber tracking is performed by using FACT [26]. The workflow is the following:

(i) Import data: an automatic registration allows to place DTI data correctly according with the structural images.

(ii) Fiber tracking: to perform tractography planning 5 different geometrical shapes can be selected for fiber tracking; ellipsoid, cube, polygon, free hand, or

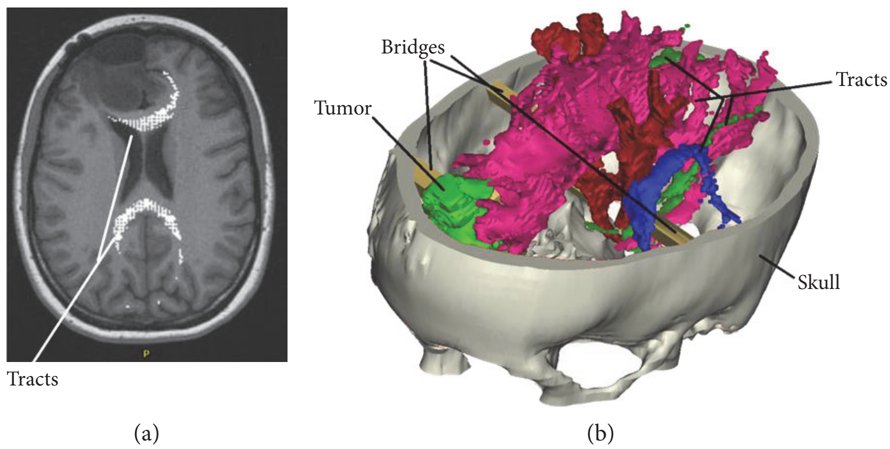


Figure 2: TRACTS superimposed on MRI structural data (a). 3D model including skull from CT, tumor from MRI, fiber tracts, and connecting bridges (b).

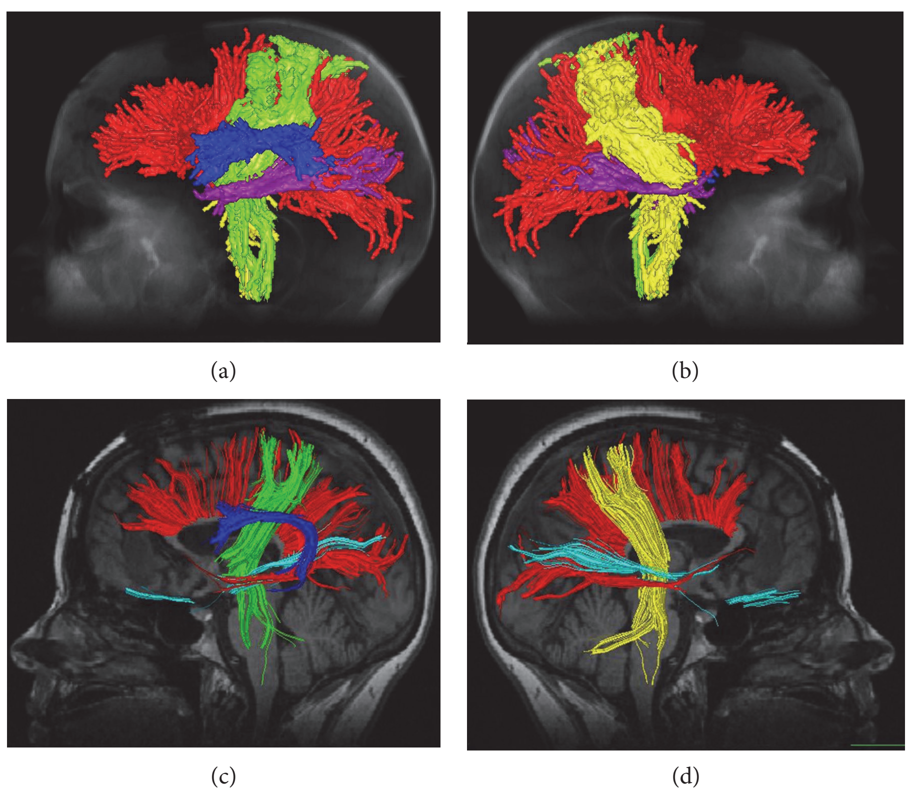
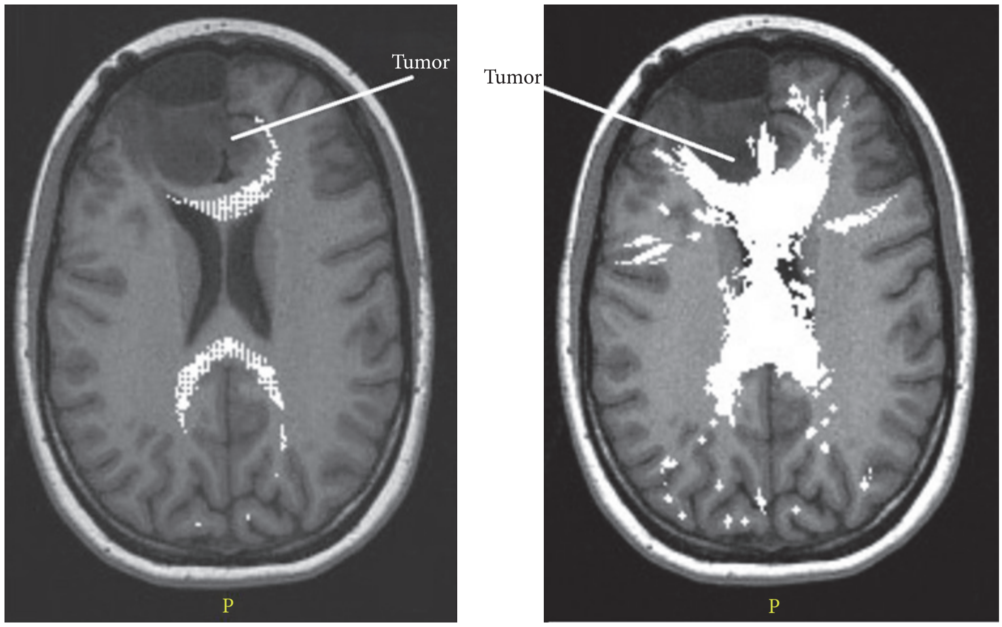


Figure 3: Tractography planning from StealthViz (a-b) and nordicBrainEx (c-d). Red color represents fiber tracts from corpus callosum, green color the tracts from motor and sensory area on the left side, and yellow the same tracts on the right side. Dark blue represents the arcuate fasciculus and purple shows the optic nerves.

scatter. These geometrical shapes are used to define volume of interest (VOI) and find the tracts of interest. On a defined VOI, three logical operators are available: (1) AND which only visualize fibers passing through that VOI, (2) OR which will only visualize fibers passing through this and any other VOIs defined, or (3) NOT that will disregard all fibers passing through that VOI. When finished tracking, one fiber, for example, corpus callosum, can be saved individually.

(iii) Export planning: results can be exported as separated files (each for every track) to the surgical navigation system or to a USB flash memory.

2.3. Quality Assessment: Anatomical Accuracy and Incorrectly Displayed Fibers. It is known that the different surgical planning software for fiber tracks may provide different results even though they are based on the same reconstruction algorithm [27]. For this reason, we performed a comparison



(a) (b)

Figure 4: Comparison between the corpus callosum tracts obtained with different software: nordicBrainEx (a) and StealthViz (b). Both (a) and (b) show the same slice.

between the software available in our institution to find the optimal one for rapid prototyping application. The assessment is based on anatomic accuracy and incorrectly displayed fibers for each fiber tract. These comparisons are done by grading the fiber-tracking results. Table 1 shows the grading for incorrectly displayed fibers and Table 2 shows the grading for anatomic accuracy. The grades vary from 1 (best) up to 4 (worst). A white matter atlas was used as a reference for the evaluation [28, 29].

2.4. Segmentation and Registration. The next step is to combine the anatomical data such as the skull and other regions of interest with the tracks from the DTI software. We use for this propose the software MIMICS [30] that is a platform for medical image processing. The process can be divided in three steps:

Step 1. CT data are imported in MIMICS and the skull bone is segmented. This operation is threshold based; where a range of HU (typically, from 600 to 2000) values are selected that allow to display the bone tissue. Next we apply *region growing* to assemble the entire connected pixel within the defined threshold in a so called *mask*. Now a 3D object can be created directly from the mask and further modification (such as opening the skull model to see inside) can be applied on the 3D object using CAD tools. Finally the 3D model can be saved as standard tessellation language (STL) file which is a format compatible with 3D Printing technologies.

Step 2. The next step is to import the tractography DICOM files to MIMICS. The 5 tracks of interest are superimposed on the anatomy (MRI data) but appearing brighter compared to those of the surrounding tissue (Figure 2(a)); therefore, the threshold-based segmentation of each tracts is easy. The five 3D objects associated with each tracks were created in the same way as described in step 1. In order to improve the quality of the 3D objects for 3D printing, we applied some morphological operations on the mask in order to smoothen details equal or below to 0.25 mm and closing distance equal to 2.5 mm (holes or gaps of 0.25 mm or less are filled). Finally, the 3D model can be saved as STL file.

Table 3: The grading results for anatomic accuracy and for incorrectly displayed fibers for both StealthViz (SV) and nordicBrainEx (BE).

Nerve tracts	Anatomic accuracy		Incorrectly displayed fibers	
	BE	SV	BE	SV
Arcuate fasciculus	3.5	1.0	3.5	3.0
Corpus callosum	2.0	1.0	2.0	1.5
Left motor and sensory tracts	2.0	1.5	2.0	1.5
Right motor and sensory tracts	2.0	1.5	2.0	1.5
Optic tracts	4.0	3.0	4.0	4.0
Total	13.5	8	13.5	12.5

Step 3. The final step is to combine Tracts, MRI, and CT data within the same 3D object. First, we imported the MRI T1-weighted images to MIMICS. Soft tissues like tumor are better visualized with MRI, and therefore, the segmentation and creation of the 3D object for this region of interest is done in this phase using the same threshold-based procedure described above. Next, we import the STL files of the skull and fiber tracts. Fiber tracts were positioned in a semiautomatic way on the 2D structural images by projecting the contours from the 3D object of the tracts. Next, we imported the

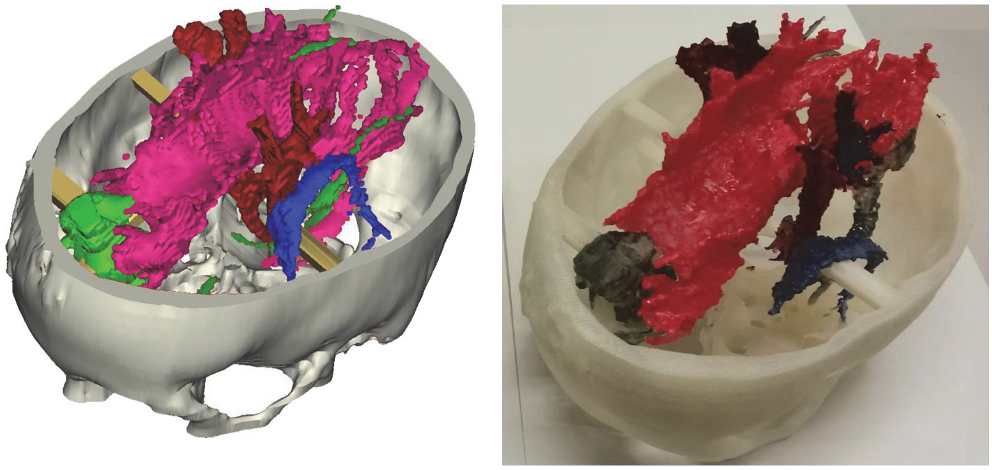


Figure 5: 3D computer model made in mimics with fiber tracts result from nordicBrainEx (a) and when it has been 3D printed and painted (b).

STL file of the skull. Since the CT and MRI data have different coordinate system, the skull 3D object is registered manually using a 3D-positioning panel. When the skull is in the right position, then the necessary connections (bridges drawn manually) between tumor, fiber tracts, and skull are built in order to create a 3D model that can be printed. Finally, the skull, tumor, fiber tracts, and the bridges are combined in one 3D object using *Boolean operations*, and the results are saved as STL (Figure 2(b)).

2.5. Navigation System and Rapid Prototyping. The 3D model of Figure 2 is finally printed using a ProJet® (3D systems, Rock Hill, USA) printer using a material called VisiJet®M3-X which is an organic colorless mixture that allow rendering of small details (mm scale). After print, the model is hardened with an infrared light.

The computer model can be exported to the surgical navigation system as DICOM set, StealthStation® [31], which works as a GPS system determining the position of surgical instruments in relation to patient images by automatically fusing CT and MRI scans. Then, a registration is done with patient anatomy; so, there is a linkage between the patient and the system. In this application, we use the 3D-printed model instead of the real patient; in this way advanced preparation of the surgery can start before the patient enters the operation theatre.

3. Results

We validate this process collaborating to a neurosurgical planning of a 29-year-old female having a low-grade glioma located on the frontal lobe.

The five fiber tracts that we focused on in this study can be seen in Figure 3, it shows the side views tractography planning from the two surgical planning software platforms: Figures 3(a) and 3(b) for StealthViz and Figures 3(c) and 3(d) for nordicBrainEx. It can be noticed that the pathways for the tracts are similar but not the identical; indeed, there are remarkable differences in thickness and ending morphology between the software platforms that may be important in relation to the pathological area of interest. In order to assess the results from the two fiber tracts planning, we use image comparison software called XERO viewer [32]; here, the fiber tracts, superimposed on the MRI data, were viewed simultaneously and visually assessed. Figure 4 shows the comparison of corpus callosum. To be noticed, the surface of corpus callosum is shown in Figure 4(b) that displays a false positive. Moreover, the fiber tract from StealthViz goes out of white matter in the brain and it is difficult to assess the exact position. Based on comparison, slice by slice and tract by tract, we assess the two tractography planning based on anatomic accuracy and incorrectly displayed fibers [28, 29]. The quantitative results are displayed in Table 3 where for this study case, the tractography plane made with nordicBrainEx has a better score and was chosen for the next step. Figure 5 shows the computer model (a) and 3D-printed model (b) resulting from the nordicBrainEx surgical planning.

DTI planning and 3D-printed models were used with the neurosurgical navigation system [12] to prepare the surgical operation where the tumor was removed from the frontal lobe. The operation was successful, and advanced planning provided with DTI planning and 3D models allowed the neurosurgeons to be better prepared during surgery.

4. Conclusion

Three-dimensional models and navigation systems for neurosurgery can be combined to improve surgical planning and surgeon training [12, 32]. The work reported herein demonstrates that preoperative planning using diffusion tensor imaging (DTI) tractography and 3D models is feasible and can be employed in the preparation of complex operations. Additionally, it is likely that this process can shorten operation times, contribute to better patient safety, and be used for training surgeons.

Even though DTI tractography is not a fully reliable method, it can still provide the neurosurgeons with an overview of fiber tract position, and it has been shown that the use of DTI improves tumor resection results and decreases postoperative deficits [14, 33]. Altogether, this work demonstrates that the reported 3D-printing process may be integrated with DTI planning and add valuable information for neurosurgical planning—especially in association with surgical navigation systems.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the Icelandic Innovation Fund RANNIS, the company Össur, and the University Hospital Landspitali, for providing financial and technical support to this project.

References

[1] T. Akiba, T. Morikawa, and T. Ohki, "Simulation of thoracoscopic surgery using 3-dimensional tailor-made virtual lung," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 143, no. 5, pp. 1232–1234, 2012.

[2] T. Akiba, T. Morikawa, and T. Ohki, "Thoracoscopic lung segmentectomy simulated by a tailor-made virtual lung: computed bronchography and angiography," *The Thoracic and Cardiovascular Surgeon*, vol. 61, no. 2, pp. 151–153, 2013.

[3] B. G. Leshnower, D. L. Miller, F. G. Fernandez, A. Pickens, and S. D. Force, "Video-assisted thoracoscopic surgery segmentectomy: a safe and effective procedure," *The Annals of Thoracic Surgery*, vol. 89, no. 5, pp. 1571–1576, 2010.

[4] P. Gargiulo and G. Á. Björnsson, "Integrated medical modeling service to optimize planning for mandibular distraction osteogenesis and maxillofacial surgeries: 5 years' experience," *Anaplastology*, vol. 2, no. 121, pp. 2161–1173, 2013.

[5] K. Watabnabe, K. Uese, O. Higuchi et al., "Three-dimensional computed tomographic findings of bilateral tracheal bronchus," *Pediatric Cardiology*, vol. 30, no. 1, pp. 87-88, 2009.

[6] Y. M. Zhong, R. B. Jaffe, M. Zhu, W. Gao, A. M. Sun, and Q. Wang, "CT assessment of tracheobronchial anomaly in left pulmonary artery sling," *Pediatric Radiology*, vol. 40, no. 11, pp. 1755–1762, 2010.

[7] R. Read, J. S. Cyr, J. Marek, G. Whitman, and A. Hopeman, "Bronchial anomaly of the right upper lobe," *The Annals of Thoracic Surgery*, vol. 50, no. 6, pp. 980-981, 1990.

[8] H. Kodama, "Automatic method for fabricating a three-dimensional plastic model with photo-hardening polymer," *Review of Scientific Instruments*, vol. 52, no. 11, pp. 1770–1773, 1981.

[9] H. Anderl, D. Zur Nedden, W. Mu et al., "CT-guided stereolithography as a new tool in craniofacial surgery," *British Journal of Plastic Surgery*, vol. 47, no. 1, pp. 60–64, 1994.

[10] P. A. Webb, "A review of rapid prototyping (RP) techniques in the medical and biomedical sector," *Journal of Medical Engineering & Technology*, vol. 24, no. 4, pp. 149–153, 2000.

[11] C. Ugo, K. J. Edmunds, and P. Gargiulo, "3D false color computed tomography for diagnosis and follow-up of permanent denervated human muscles submitted to home-based functional electrical stimulation," *European Journal of Translational Myology*, vol. 25, no. 2, p. 5133, 2015.

[12] P. Gargiulo, S. B. Sævarsdóttir, B. Baldvinsdóttir, and I. H. Ólafsson, "Use of 3 dimensional models and navigation system in preparation for brain surgery," *Biomedizinische Technik*, vol. 58, no. 1, p. 1, 2013.

[13] H. Halldórsson, A. J. Arason, M. Sigurðardóttir et al., "Recurrent tumor growth in the right maxilla–a rare case of an inflammatory myofibroblastic tumor," *Laeknabladid*, vol. 101, no. 1, pp. 19–23, 2015.

[14] D. Leclercq, C. Delmaire, N. Menjot de Champfleur, J. Chiras, and S. Lehéricy, "Diffusion tractography: methods, validation and applications in patients with neurosurgical lesions," *Neurosurgery Clinics of North America*, vol. 22, no. 2, pp. 253–268, 2011.

[15] K. Kośla, B. Bryszewski, D. Jaskólski, N. Błasiak-Kołacińska, L. Stefańczyk, and A. Majos, "Reorganization of language areas in patient with a frontal lobe low grade glioma–fMRI case study," *Polish Journal of Radiology*, vol. 80, no. 1, p. 290, 2015.

[16] D. Frey, V. Strack, E. Wiener, D. Jussen, P. Vajkoczy, and T. Picht, "A new approach for corticospinal tract reconstruction based on navigated transcranial stimulation and standardized fractional anisotropy values," *NeuroImage*, vol. 62, no. 3, pp. 1600–1609, 2012.

[17] V. A. Coenen, T. Krings, L. Mayfrank et al., "Three-dimensional visualization of the pyramidal tract in a neuronavigation system during brain tumor surgery: first experiences and technical note," *Neurosurgery*, vol. 49, no. 1, pp. 86–93, 2001.

[18] Medronic, "StealthViz with StealthDTI application reference guide," Medronic, 2008–2011.

[19] NordicNeuroLab AS, "nordicBrainEx Tutorial - DTI Module," NordicNeuroLab AS, 2015.

[20] A. Fitsiori, D. Nguyen, A. Karentzos, J. Delavelle, and M. I. Vargas, "The corpus callosum: white matter or terra incognita," *The British Journal of Radiology*, vol. 84, no. 997, pp. 5–18, 2011.

[21] Bioon.com, "Basic motor pathway," February 2017, http://www.bioon.com/bioline/neurosci/course/basmot.html.

[22] Study.com, "Sensory cortex: definition & function - video & lesson transcript," February 2017, http://study.com/academy/lesson/sensory-cortex-definition-function.html.

[23] Healthline.com, "Optic nerve function, anatomy & definition | body maps," February 2017, http://www.healthline.com/human-body-maps/optic-nerve.

[24] A. D. Friederici, "Pathways to language: fiber tracts in the human brain," *Trends in Cognitive Sciences*, vol. 13, no. 4, pp. 175–181, 2009.

[25] F.-C. Yeh, T. D. Verstynen, Y. Wang, J. C. Fernández-Miranda, and W.-Y. I. Tseng, "Deterministic diffusion fiber tracking improved by quantitative anisotropy," *PloS One*, vol. 8, no. 11, article e80713, 2013.

[26] G. C. Feigl, W. Hiergeist, C. Fellner et al., "Magnetic resonance imaging diffusion tensor tractography: evaluation of anatomic accuracy of different fiber tracking software packages," *World Neurosurgery*, vol. 81, no. 1, pp. 144–150, 2014.

[27] M. Catani and M. Thiebaut de Schotten, "A diffusion tensor imaging tractography atlas for virtual in vivo dissections," *Cortex*, vol. 44, no. 8, pp. 1105–1132, 2008.

[28] K. Oishi, A. V. Faria, P. C. M. van Zijl, and S. Mori, *MRI Atlas of Human White Matter*, Elsevier Academic Press, China, 2010.

[29] MIMICS software, February 2017, http://biomedical.materialise.com/mimics.

[30] Xero viewer software, February 2017, http://agfahealthcare.com/he/usa/en/binaries/XERO%20sheet%202013_tcm561-113934.pdf.

[31] Medtronic, http://www.medtronic.com/us-en/healthcare-professionals/products/neurological/surgical-navigation-systems/stealthstation.html.

[32] P. Gargiulo, "3D modelling and monitoring of denervated muscle under functional electrical stimulation treatment and associated bone structural change," *European Journal of Translational Myology/Basic Applied Myology*, vol. 21, no. 1, pp. 31–94, 2011.

[33] O. Ciccarelli, M. Catani, H. Johansen-Berg, C. Clark, and A. Thompson, "Diffusion-based tractography in neurological disorders: concepts, applications, and future developments," *Lancet Neurology*, vol. 7, no. 8, pp. 715–727, 2008.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 8264071, 11 pages
https://doi.org/10.1155/2017/8264071

Research Article

Fall Prevention Shoes Using Camera-Based Line-Laser Obstacle Detection System

Tzung-Han Lin,[1] Chi-Yun Yang,[2] and Wen-Pin Shih[2]

[1]*Graduate Institute of Color and Illumination Technology, National Taiwan University of Science and Technology, No. 43, Sec. 4, Keelung Rd., Taipei 10607, Taiwan*
[2]*National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan*

Correspondence should be addressed to Tzung-Han Lin; thl@mail.ntust.edu.tw

Received 9 February 2017; Accepted 20 April 2017; Published 15 May 2017

Academic Editor: Junfeng Gao

Fall prevention is an important issue particularly for the elderly. This paper proposes a camera-based line-laser obstacle detection system to prevent falls in the indoor environment. When obstacles are detected, the system will emit alarm messages to catch the attention of the user. Because the elderly spend a lot of their time at home, the proposed line-laser obstacle detection system is designed mainly for indoor applications. Our obstacle detection system casts a laser line, which passes through a horizontal plane and has a specific height to the ground. A camera, whose optical axis has a specific inclined angle to the plane, will observe the laser pattern to obtain the potential obstacles. Based on this configuration, the distance between the obstacles and the system can be further determined by a perspective transformation called homography. After conducting the experiments, critical parameters of the algorithms can be determined, and the detected obstacles can be classified into different levels of danger, causing the system to send different alarm messages.

1. Introduction

When the elderly fall, it can harm their bodies and has serious negative mental impacts on them. In the United States, about 30% of adults above 65 years of age suffer a fall annually [1]. Besides, they bring about a tremendous medical expense in the society. For example, the total cost of fall-related hospitalizations in Texas in 2011 was 3.9 million US dollars. For people aged above 65, the total cost was 2.5 million. As the proportion of the elderly in societies grows, the chances of falls will become larger.

There are many risk factors that can cause the elderly to fall, and they can be roughly classified into three categories: intrinsic factors, extrinsic factors, and exposure to risk [2]. Among the intrinsic factors, different kinds of diseases and reduced physical abilities may increase the probability of falls. As for the extrinsic factors, some studies show that 30% of falls of the elderly are due to accidents and environment-related risk factors [3]. The extrinsic factors include environmental hazards, that is, poor lighting, slippery floors, uneven surfaces, obstacles, unsuitable footwear and clothing, and inappropriate walking aids or assistive devices.

With respect to exposure to risks, Graafmans et al. suggest that the most inactive and the most active people have the highest risk of falls [4]. The most inactive among the elderly may have worse physical capabilities. On the other hands, those who indulge in physical activities can maintain their neuromuscular functioning, which is necessary to keep a balance and to react to falls. However, physical activities also increase greater exposure to environmental risks.

Researchers have worked for a long time to develop various kinds of fall prevention and fall detection systems [5, 6]. Until now, most fall detection systems recognized fall events after they happened. Therefore, a line-laser obstacle detection system that detects environmental obstacles to prevent falls in advance is provided in this paper. The system is installed on the toe end of the shoes, and when the obstacles are recognized, the system will send alarm messages to catch the attention of the elders, as shown in Figure 1. In this manner, the elderly can notice the obstacles that may cause them to fall in advance, thereby reducing the risk of falls.

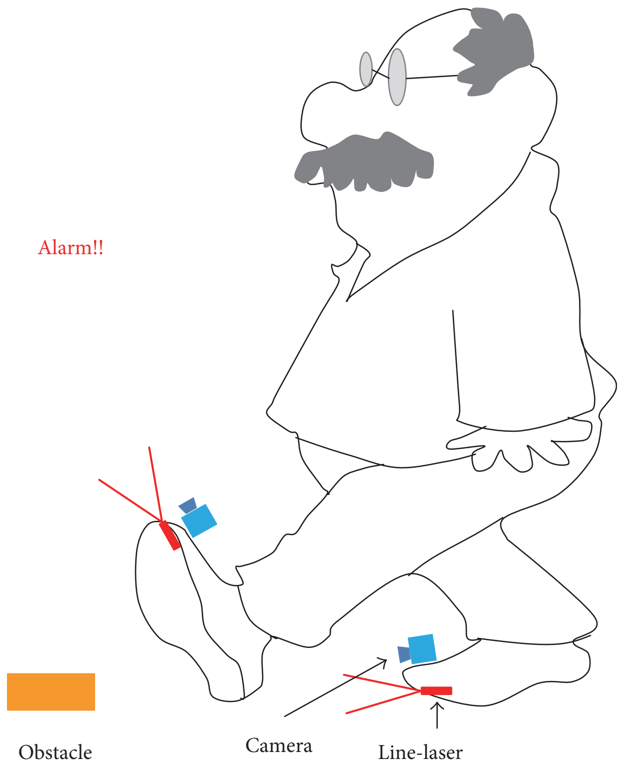


Figure 1: Line-laser obstacle detection system sends alarm messages to catch the attention of the elderly when they encounter dangerous obstacles.

As a result, physical damages and a negative impact on the mental state of the older people can be prevented and the medical costs of the whole society can be saved to a large extent [7].

2. Related Work

In order to detect the environmental hazards that may cause falls, related obstacle detection methods and technologies were deeply investigated. Traditionally, obstacle detection has been vital for many mobile robot applications [8]. Popular range-based sensors include ultrasonic sensors, laser range finders, radar, and stereo vision. Each kind of range-based sensor has its limitations. Ultrasonic sensors are cheap but suffer from specular reflection problems. Laser range sensors have higher resolution, but they are more complex and more expensive than ultrasonic sensors. Stereo vision is computationally expensive and difficult to be a practical solution.

Engineers have developed methods and various kinds of systems for obstacle detection based on the abovementioned range sensors. Manduchi et al. implemented a color stereo camera and a single axis radar for cross-country autonomous systems [9]. Zingg et al. presented an approach for wall collision avoidance using a depth map based on optical flow from on board camera images [10]. Oniga and Nedevschi provided a method to classify road surface, traffic isle, and obstacle detection using rectangular digital elevation map from dense stereo data [11]. Zhang et al. developed a novel algorithm for on-road obstacle detection based on stereo cameras [12]. The proposed algorithm in this paper significantly reduces the complexity disparity calculations involved when using a stereo vision technique. In addition, Batavia and Singh developed an obstacle detection methodology, which combines two algorithms: adaptive color segmentation and stereo-based color homography [13]. This algorithm is particularly suited for environments where the terrain is relatively flat and of roughly the same color.

With respect to laser range finders, Fu et al. designed an integrated triangulation laser scanner for obstacle detection installed on miniature mobile robots [14]. The basic components of the triangular laser system are composed of a laser emitter and a camera; therefore, the system becomes smaller and less power demanding. As a result, it is possible to integrate the triangular laser scanners on microrobots.

Stereo vision is, however, computationally expensive. On the contrary, a camera-based line-laser technique that needs much less computation amounts is proposed to detect obstacles in this paper which is a part of Yang's thesis [15]. Because of its lesser computational amount, the extracted line-laser pattern has high potential to be implemented on embedded systems. In addition, the simple components reduce the costs of the overall system. Therefore, a complete integration of the system on shoes for the purpose of fall prevention can be achieved.

3. Method

3.1. System Configuration. In our system configuration, a line-laser is mounted on the side of the shoes and an RGB camera is fixed but tilted down on the top side of the shoes [16]. The relative position between the line-laser and the camera is extremely rigid to deliver a consistent obstacle detection result. To have a good depth resolution, the distance between these two components should be as large as possible. The tilted angle of the camera can be adjusted according to the detection region in our prototype. Because the average step length of older people is around 0.808 m [17, 18], the detection region of our system is 0.5–1.0 m.

In our prototype, a Logitech C310 webcam that operates at 29 frames per second under a 640×480 resolution and a 405 nm wavelength laser are used. Both of them are mounted rigidly to have consistent calibration parameters. Moreover, to suppress the noise interference, a blue glass paper is used as a band-pass filter to resist the unnecessary light from the environment, as shown in Figure 2.

3.2. Software Framework. Our algorithm flowchart is shown in Figure 3. Initially, the camera continuously acquires an image and then the obtained image is compared with the previous one. If the difference between two successive frames is very small, we assume that the users step on the ground. Besides, the moment the users raise the foot to the max height during a gait cycle also has the same effect. In both the above cases, the event for determining the obstacles will be triggered. Otherwise, the program will

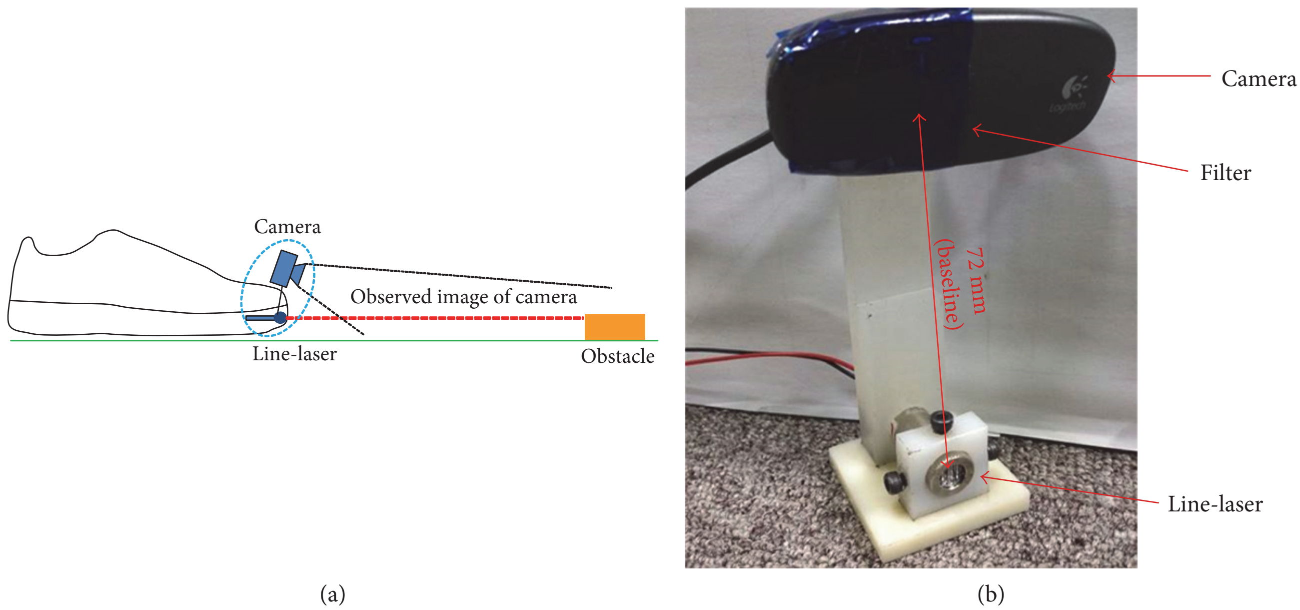


Figure 2: (a) System configuration of line-laser obstacle detection system integrated on shoes and (b) prototype of the proposed system.

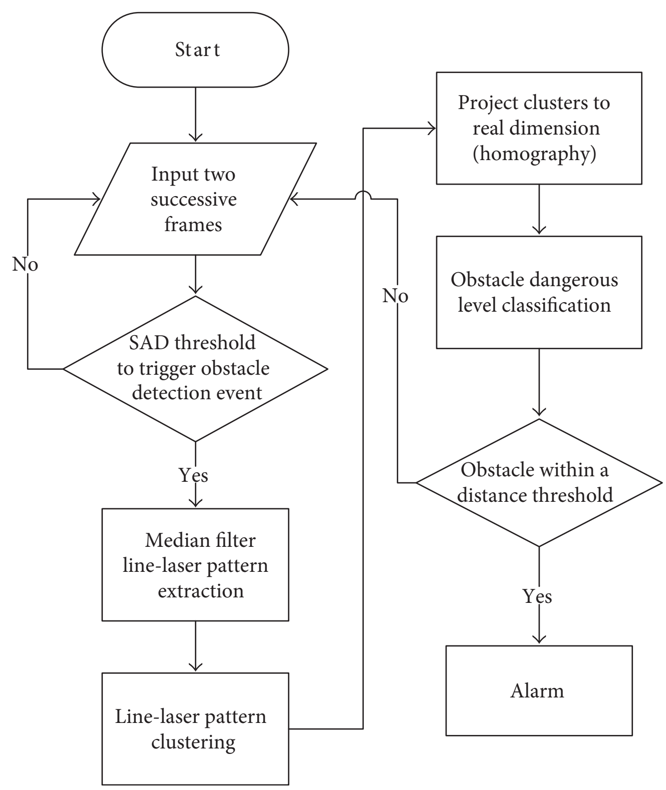


Figure 3: Software framework of line-laser obstacle detection system.

only continuously calculate the difference between two incoming frames.

Once the obstacle detection event is triggered, a series of line-laser pattern extraction steps will be executed. First, the median filter is applied in order to suppress noises. In addition, an intensity threshold is used to identify the pixels where the laser light is. Then, the max value of each column of the image is extracted and considered as a potential laser pattern. After that, a segmentation method is applied to classify every laser pixel into several clusters that may denote the obstacles. Some clusters that do not represent obstacles are removed, and the clusters that denote the same obstacles together are merged. Subsequently, the real depths and widths of the obstacles are obtained by homography transformation. Finally, the system will send alarm messages according to the dangerous levels of the obstacles classified by their

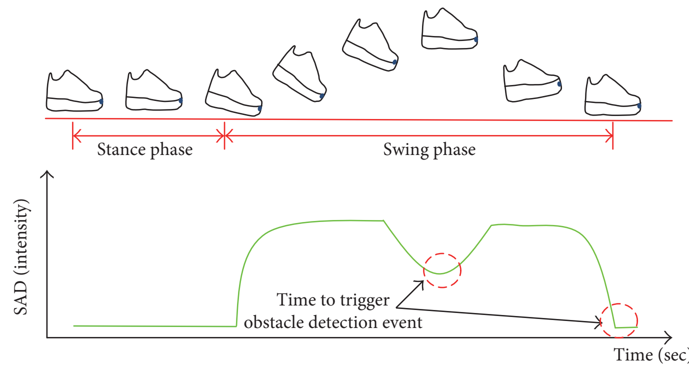


Figure 4: System trigger time determined by the SAD value.

widths and depths. Each step in the software framework is illustrated in detail as follows.

3.3. Sum of Absolute Difference (SAD) Threshold to Trigger the Obstacle Detection Event. In order to detect the obstacles, the obstacle detection event is triggered when a user steps on the ground. Besides, to detect higher obstacles, the event is triggered when users raise their feet at the max height during a gait cycle as well. In both cases, their feet are roughly horizontal to the ground. Therefore, the system is suitable to detect the obstacles in front of the users at these moments. Besides, this strategy will save computation consumption at other times during a gait cycle.

The difference between the two frames that are captured by the camera is utilized to identify when the users either step on the ground or raise their feet at the max height. This concept was first proposed by Fitzpatrick and Kemp [19]. In our method, the normalized sum of absolute difference (SAD) value is utilized to compute the difference between two successive frames as in

$$\mathrm{SAD} = \frac{\sum_{u=0}^{N_{\mathrm{rows}}} \sum_{v=0}^{N_{\mathrm{cols}}} \left| I_p(u,v) - I_l(u,v) \right|}{N_{\mathrm{rows}} N_{\mathrm{cols}}}, \qquad (1)$$

where u and v denote the pixel's coordinates on row and column directions. N_{cols} and N_{rows} are the pixel numbers of the image width and height. I_p and I_l represent the pixel intensity of the current frame and the previous frame, respectively. In general, the brightness of a color pixel comes from a linear combination of R, G, and B components. The intensity we used is only the B component. However, the intensity value of the camera highly depends on the environment illumination. Therefore, we turn off the auto white balance function and lock the exposure time for the camera.

In a gait, there are two phases: stance phase and swing phase. Usually, the SAD value between frames during a stance phase should be very small. Similarly, the SAD value will be small when the feet are at the max height. Once the users start to move their feet forward, the SAD value will become very large in contrast to the stance phase. In our experiment, the relatively small SAD value happens at the middle and at the end of every swing phase as shown in Figure 4. Therefore, the obstacle detection will be triggered when the SAD value is small.

3.4. Line-Laser Pattern Extraction. Once the obstacle detection event is triggered, a median filter is then applied to suppress the bad influences coming from the environmental noise [20]. The median filter with a specific window size sweeps across the whole image, and the intensity value of the middle pixel in the window will be replaced by the median of pixel values in the window. Therefore, it can maintain the original structure well, particularly for edge features.

After applying the median filter, an intensity threshold is set to separate the line-laser pattern from the background. At this step, any pixel with an intensity below the threshold is recognized as background or noise. On the other hand, if the intensity of a pixel exceeds the intensity threshold, it will be considered as a candidate pixel on the obstacles.

Since our system is mainly designed for indoor application, the intensity distribution of different obstacles that are detected by the system is needed to be investigated. Therefore, twelve common construction materials are surveyed, as arranged in Figure 5. After implementing detection experiment of each material, the intensity distribution data are depicted as shown in Figure 6, except that a mirror will directly reflect the laser light; therefore, it cannot be detected successfully by our system. In our application, the intensity threshold 14 is selected because 95% of the laser light among the experimental cases will pass the intensity threshold. On the contrary, if the intensity of any pixel is below the intensity threshold 14, it is recognized as a part of the background.

Subsequently, a line-laser pattern is extracted by the following steps. The pixel having the max intensity value in each column is collected to be the line-laser pattern pixel. However, to increase the performance, the image is initially rotated 90° for processing and then rotated back. In addition, because of the need to reduce the impact of noise, the average of the pixel intensities within a window is searched for each column instead, as indicated in Figure 7(a). If the average value exceeds a specific threshold, the centroid of the window will be considered as the location of line-laser pattern in this column.

After the extraction of line-laser pattern, the extracted data are stored and processed again for obstacle clustering. Because of unstable line-laser intensities from the camera and disturbances from the environmental noise, the

Figure 5: Common construction materials used indoors.

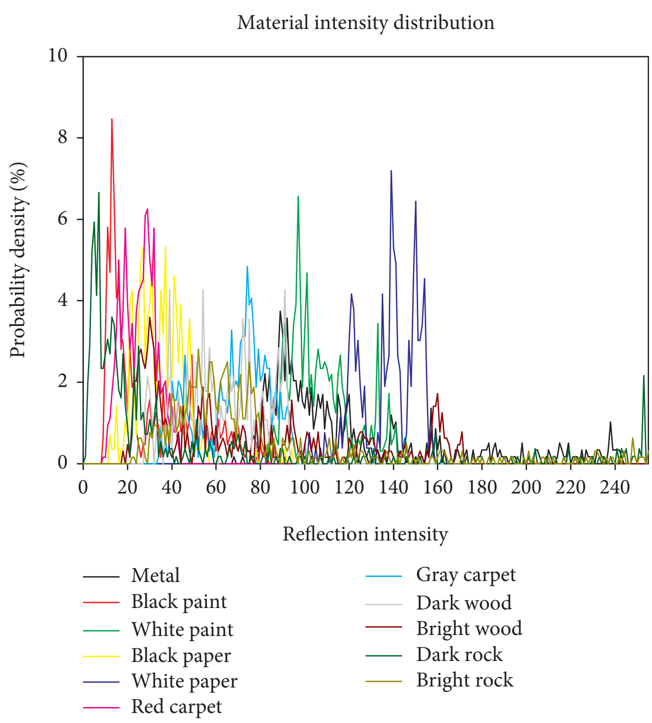


Figure 6: Reflection intensity distribution of common construction materials.

extracted line-laser pattern may suffer negative influence. In Figure 7(b), the example shows several noise groups and nonclustered obstacles. In these cases, noises will be rejected while the neighboring pixels are merged. In addition, a candidate obstacle will propagate due to the merging of neighboring clusters.

3.5. Line-Laser Pattern Clustering. The goal of our system is to recognize how far and how large the potential obstacles are. Therefore, a segmentation algorithm is needed after executing the line-laser pattern extraction procedure. This algorithm classifies each pixel on the line-laser pattern into several clusters that are likely denoting the obstacles.

We assume all pixels of an obstacle will form only a continuous segment. In condition, the distance deviation of the current cluster in the image domain will not exceed a specific threshold T as illustrated in Figure 8. Thus, a region-growing procedure is applied to one of unclassified pixels of the line-laser pattern. Therefore, several clusters that denote the obstacles are obtained.

A suitable threshold T in an indoor environment is critical. The experiment to determine T is described in a later

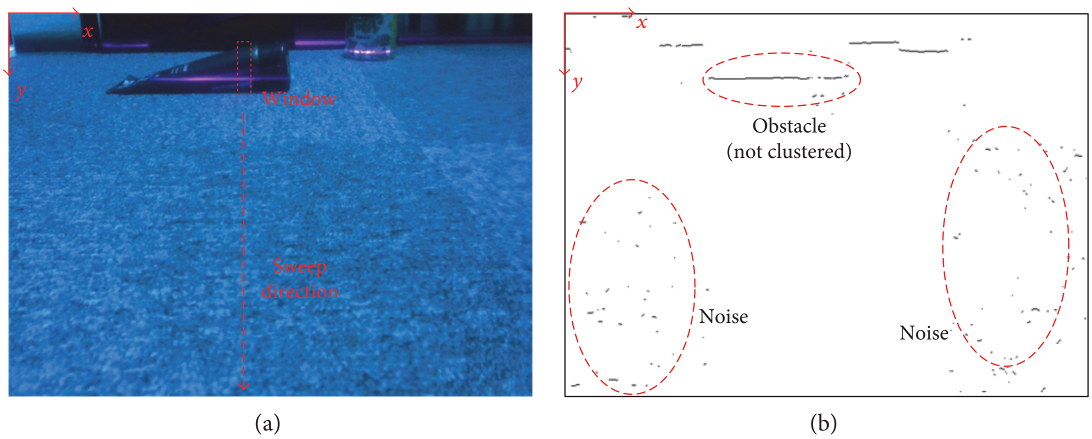


Figure 7: (a) Max average value of pixels in a window is searched to be the extracted line-laser pixel. (b) Line-laser pattern extraction result. The clusters in the red circle, which denote the same obstacle, break into parts, and some small clusters from the noise exist in the image.

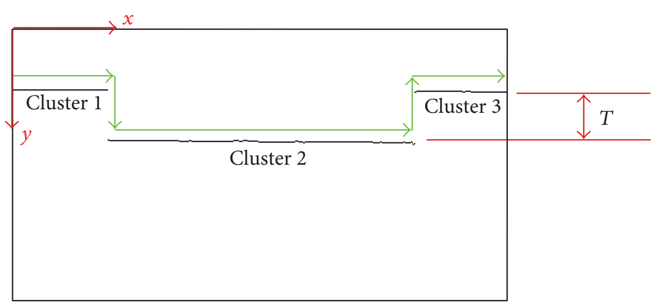


Figure 8: Search for line-laser pattern pixel by pixel for segmentation. If the y coordinate of the present pixel deviates from the last pixel and exceeds deviation threshold T, the present pixel is classified into another cluster.

section. Based on the discussion of the deviation threshold T, our deviation threshold for line segmentation is two pixels. Nevertheless, some clusters that belong to the same obstacle break apart, and they should be merged. We first eliminate small clusters that have very few pixels, less than six, and then merge the residuals by a deviation threshold.

In order to eliminate the small clusters from the environmental noise, the width of the segments, which come from the noise of images, are counted. The statistical result indicates that the widths of 95% clusters of noises are below six pixels. Therefore, clusters with widths less than or equal to six pixels are recognized as noises. Furthermore, clusters with less than six pixels will have a 0.45 cm of width in the middle of our working region. If the horizontal distance between two neighboring clusters is less than the width of a foot, people may still kick on the gap between two obstacles and then trip over. Therefore, on the condition that the y coordinate deviation on extracted line-laser image between two clusters is within the deviation threshold T, and the real world distance of the gap between two clusters is less than the width of a foot (10 cm), these two clusters should be merged. An example is illustrated in Figure 9(a). The merging step is taken from the leftmost cluster to the right. The result after executing noise elimination and the merging steps is shown in Figure 9(b).

3.6. Homography Transformation. After executing the line-laser pattern clustering, the physical distance is needed to be determined by homography transformation. The calibration for homography transformation in our paper utilizes a checkerboard. The transformation denotes the relationship between the image coordinate and the real world coordinate of the checker grid, as shown in Figure 10. Since the detection region of our prototype is 0.5–1.0 m, the obstacle detection does not need very accurate distance estimation. Therefore, the lens distortion of the camera in our prototype is negligible.

A coordinate $[x, y, 1]^T$ on the image plane can be mapped into the real world coordinate $[x_1', x_2', x_3']^T$, which indicates a homogeneous coordinate after applying a 3×3 homography matrix, as (2) [21]. The real world coordinate can be converted into a real dimension by (3).

$$\begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{2}$$

$$(x', y') = \left(\frac{x_1'}{x_3'}, \frac{x_2'}{x_3'} \right). \tag{3}$$

By rewriting (2) and (3), eight unknowns in the homography matrix become (4), where $h_{33} = 1$. Because there are eight

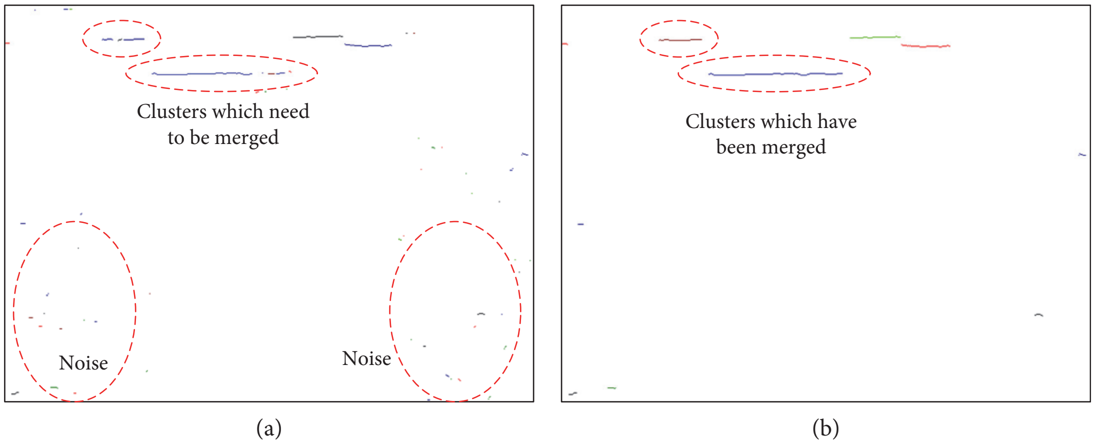


Figure 9: (a) Result of line-laser pattern clustering. Some residual noise needs to be further eliminated. Besides, clusters that belong to the same obstacle should be merged. (b) The result after executing noise illumination and merging.

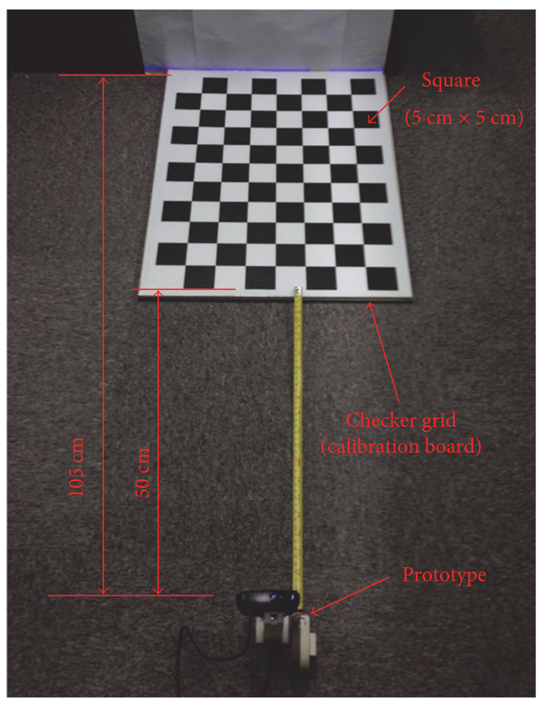


Figure 10: Calibration setup of our prototype.

unknowns to be solved, we need at least four corresponding points to solve the homogeneous matrix. In the calibration procedure, the calibration board is put in front of our prototype. In addition, the line-laser roughly passes through the surface of the calibration board. After taking a picture for the checker grid, four points on the image and their corresponding cross corners on the checker grid are utilized to determine the homography matrix, as shown in Figure 11. For example, four cross corners in real coordinate with centimeter will be (5, 5), (35, 5), (35, 50), and (5, 50) in sequence. By solving (4), eight unknowns representing a homography matrix are obtained. In practice, the real world coordinate should be shifted again due to a translation in the laser's position.

$$\begin{bmatrix} x & y & 1 & 0 & 0 & 0 & -x'x & -x'y \\ 0 & 0 & 0 & x & y & 1 & -y'x & -y'y \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix}. \tag{4}$$

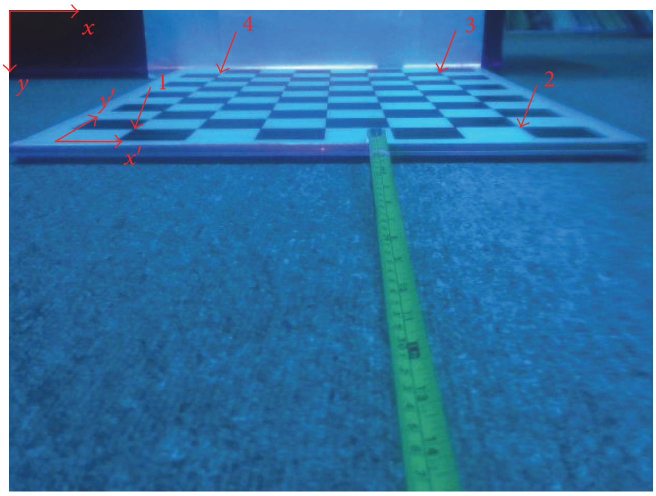


FIGURE 11: Four cross corners in real world coordinates and their corresponding points in the image coordinates are utilized to determine the homography transformation between two coordinates.

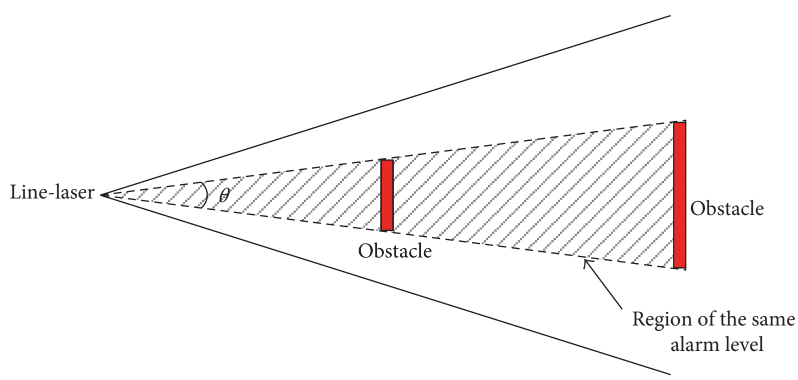


FIGURE 12: Obstacle alarm level of each obstacle can be classified by the angle θ in this figure. The two obstacles are both in the triangular shaded region, and thus, their alarm levels are the same.

3.7. Obstacle Alarm Level Classification. As long as the width and the depth of each obstacle can be calculated, the obstacle alarm level of each obstacle can be classified. Since the size and distance may affect the dangerous level, we use a coverage angle θ to a normalized factor for alarm as shown in Figure 12. The two obstacles in Figure 12 are of the same coverage angle to the laser position, and thus, their alarm levels are considered to be the same. In our proposed system, alarm level classification is used because residual small clusters may cause the system to send false positive alarm messages on the condition that the system just determines whether dangerous obstacles exist.

The system distinguishes the obstacles into four alarm levels. Alarm level 1 means the obstacle is the most dangerous, causing the system to send the most urgent and loud alarm messages. On the other hand, when alarm level of the obstacles becomes 2 or 3, the alarm message will become weaker. Last, if the alarm level of the obstacles is 4, the system will not react and send any alarm messages.

4. Result and Performance Evaluation

4.1. SAD Threshold Determination. In order to determine a proper SAD threshold for triggering the obstacle detection event, the camera continues recording SAD values during a gait cycle. A person wore the shoes and then walked around for a period of time. The collected data are shown in Figure 13. It is clear that a periodical shape comes out. After collecting the data, the SAD threshold, which represents the transition SAD value of the stance phase and swing phase, is obtained by averaging the SAD values of all the frames in one gait. A threshold equaling to 15.8 is therefore obtained. In Figure 13, the SAD value in the stance is stable and is as low as 8.2. To define a strict threshold, a threshold value of 12 is finally determined.

Apart from the moment a user steps on the ground, the obstacle detection event should be triggered when people raise their foot at max height to detect higher obstacles. However, a SAD threshold to determine the moment in swing phase is difficult to be identified because of the unsteady

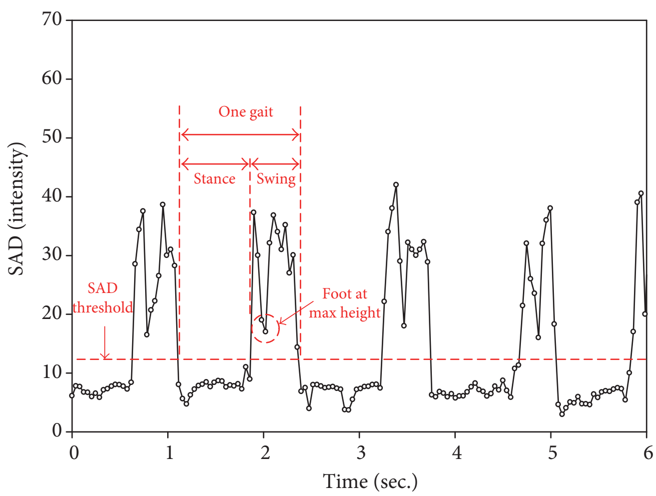


Figure 13: SAD values of several gaits. The threshold is determined below the average of SAD values in one gait.

characteristics of different steps. Besides, it may vary in the gait behavior. Therefore, a reasonable definition for the moment when the elders' foot at max height highly depends on the collection of data. In our experiment, the moment is roughly at one third of a gait cycle, which has a relative small SAD value within the 3rd–7th frames during the swing phase. In practice, these five frames are temporarily stored, and the smallest SAD value is then selected as the moment for triggering obstacle detection event.

4.2. Deviation Threshold T. In obstacle detection event, we use a deviation threshold T for rejecting neighboring pixels that do not belong to the current segment. The deviation threshold value T for the line segmentation is critical. If the deviation threshold T is set at a value too small, one cluster may be classified into many different smaller clusters, and this may cause the original cluster to be neglected by our system. On the other hand, if the deviation threshold is chosen to be too large, some clusters that belong to different obstacles may be grouped together. As a result, the system will over emit alert signals to the users.

In our system, a line-laser casts light on the obstacles and one camera observes the line patterns. Therefore, the cast line becomes broader and stronger at a smaller distance due to the perspective effect. However, after analyzing the line-laser pattern at each distance, the deviation threshold $T = 1$ is the same for all distances in the detection region 0.5–1.0 m. In other words, a suitable deviation threshold T will not change with detection distance when the line-laser obstacle detection system detects the same obstacle.

On the other hand, the deviation of pixels of the same cluster varies and depends on the reflectivity of obstacles. Therefore, an experiment is designed to find out a suitable deviation threshold T of common construction materials that are usually used indoors. Twelve test samples are considered as shown in Figure 5. Because a suitable deviation threshold T will not vary with the detection distance, all the experiments were carried out at a distance of 100 cm. After the experiment, a conclusion is drawn that except the dark rock (sample number 11) that needs the deviation threshold T to be 2 pixels for correct segmentation and that the mirror (sample number 2) cannot be detected directly because of its high reflectivity, the suitable deviation threshold for the rest of the samples is 1 pixel. However, in order to handle all the cases strictly, the proper deviation threshold T is set to be 2 pixels for the correct operation of our line-laser obstacle detection system.

Table 1: Distance estimation.

Ground truth distance (cm)	Estimated distance (cm)	Error (cm)	Error (%)
50	51.7	1.7	3.4
60	61.1	1.1	1.8
70	70.4	0.4	0.5
80	79.8	−0.2	−0.3
90	88.9	1.1	−1.2
100	99.5	−0.5	−0.5

4.3. Distance Estimation by Homography. Furthermore, an experiment is carried out to validate the precision of the distance estimation by homography. The line-laser obstacle detection system is mounted on a linear slider rigidly, and then the linear slider is used to adjust the distance between our system and a flat wall. The distance computation result by homography mapping will be compared to the standard distance measurement by the slider. From the validation experiment, the measurement error at 80 cm, the middle of the working distance, is less than 0.3 cm. At the nearest working distance, say 50 cm, the measured error is 1.7 cm, as shown in Table 1. This error value is acceptable compared to the step lengths of the elderly.

4.4. Obstacle Alarm Level Classification Results. The line-laser obstacle detection system classifies the obstacle alarm levels according to their angles. We define four alarm levels. When the obstacle coverage angle $\theta > 7.8^\circ$, the alarm level is 1. That represents that an obstacle with a 10 cm width, as well as the foot width of an adult, is located as far at 75 cm in front of the user's shoes. The alarm level 2 is assigned when the angle θ is between 3.8° and 7.8°. Similarly, alarm level 3 is raised when the angle θ is between 1.5° and 3.8°. If the angle θ is less than 1.5°, which indicates that the obstacle's width is as small as 2 cm in the middle of the working distance, alarm level 4 is obtained and no alarm signal will be sent.

The system performance is tested by detecting real obstacles, which may occur regularly indoors, as shown in Figure 14. The experiment results prove that the system can work well to compute the widths and the distances of obstacles and then determine their alarm messages. In Figure 14(a), four cases are included and the highest priority, say level 1, will be finally sent. Based on this strategy, the probability of false positive errors will be low, and the alarm messages can be sent out according to the dangerous levels of encountered obstacles.

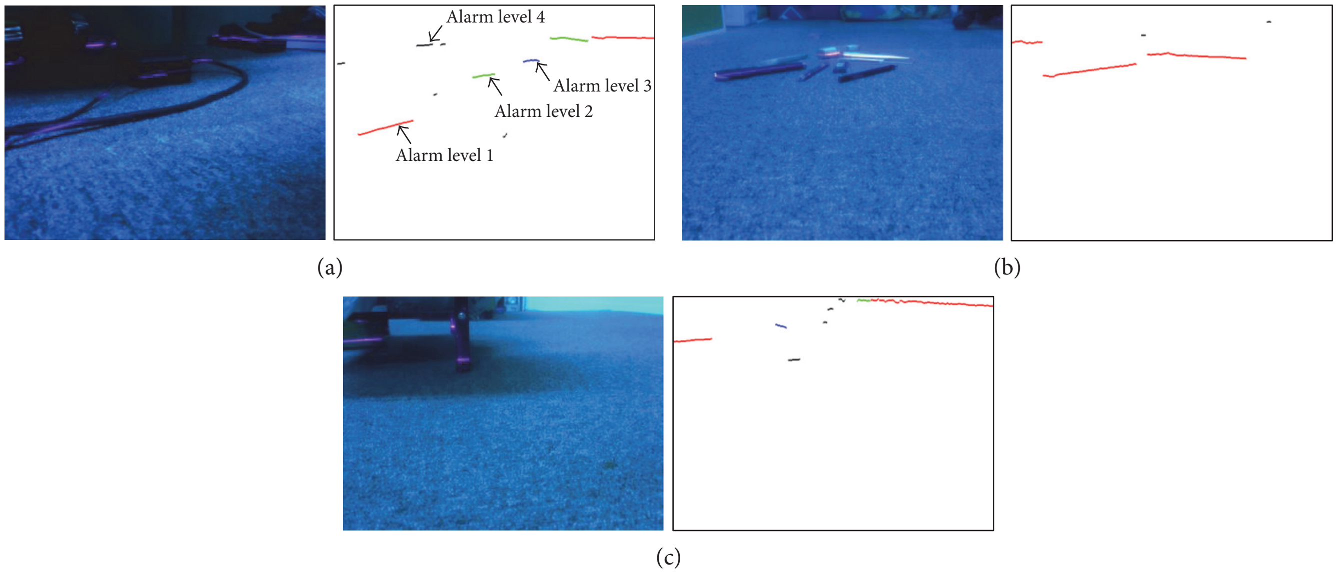


Figure 14: (a) System detects a cable on the ground. The red clusters represent alarm level 1, the green clusters represent alarm level 2, and the blue clusters represent alarm level 3. The black clusters denote alarm level 4. (b) System detects scatterings on the ground. (c) The system detects a corner of a table.

4.5. Limitations of the Proposed Method. With regard to the limitations of our method, the camera may suffer from the drawback of its low dynamic range. Besides the height of detectable, the obstacles must be larger than the plane level of the line-laser. Nevertheless, the feature of fall prevention is carried out by strategically sending an alarm message to the users.

5. Conclusion

In this paper, a camera-based line-laser obstacle detection system is proposed for designing fall prevention shoes of users. The system simply consists of an RGB camera, a filter, and a line-laser, so it is suitable to be installed on customer wearable devices, and the overall costs of the products are acceptable compared to shoes. We successfully verified the algorithms, including SAD threshold to trigger the obstacle detection event, line-laser pattern segmentation, homography transformation, and obstacle dangerous level classification. Finally, a prototype for the prevention of falls of the elderly was carried out.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Science Council of Taiwan under Grant no. NSC-100-2628-E-002-025-MY3.

References

[1] S. D. Towne, M. G. Ory, and M. L. Smith, "Cost of fall-related hospitalizations among older adults: environmental comparisons from the 2011 Texas hospital inpatient discharge data," *Population Health Management*, vol. 17, no. 6, pp. 351–356, 2014.

[2] C. Todd and D. Skelton, *What Are the Main Risk Factors for Falls among Older People and What are the Most Effective Interventions to Prevent These Falls?*, World Health Organization Regional Office for Europe, Copenhagen, 2004.

[3] L. Z. Rubenstein, "Falls in older people: epidemiology, risk factors, and strategies for prevention," *Age and Ageing*, vol. 35, Supplement 2, pp. ii37–ii41, 2006.

[4] W. C. Graafmans, P. Lips, G. J. Wijlhuizen, S. M. Pluijm, and L. M. Bouter, "Daily physical activity and the use of a walking aid in relation to falls in elderly people in a residential care setting," *Zeitschrift für Gerontologie und Geriatrie*, vol. 36, no. 1, pp. 23–28, 2003.

[5] X. Yu, "Approaches and principles of fall detection for elderly and patient," in *Applications and Services, in 10th International Conference on e-health Networking* of *IEEE Communications Society*, pp. 42–47, Singapore, July 2008.

[6] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: principles and approaches," *Neurocomputing*, vol. 100, pp. 144–152, 2013.

[7] S. Sadiqh, A. Reimers, R. Andersson, and L. Laflamme, "Falls and fall-related injuries among the elderly: a survey of residential-care facilities in a Swedish municipality," *Journal of Community Health*, vol. 29, no. 2, pp. 129–140, 2004.

[8] I. Ulrich and I. Nourbakhsh, "Appearance-based obstacle detection with monocular color vision," in *Proceedings of the AAAI National Conference on Artificial Intelligence*, pp. 866–871, Austin, TX, USA, July 2000.

[9] R. Manduchi, A. Castano, A. Talukder, and L. Matthies, "Obstacle detection and terrain classification for autonomous off-road navigation," *Autonomous Robots*, vol. 18, no. 1, pp. 81–102, 2005.

[10] S. Zingg, D. Scaramuzza, S. Weiss, and R. Siegwart, "MAV navigation through indoor corridors using optical flow," in *IEEE International Conference on Robotics and Automation*

(ICRA) of *IEEE Computer Society*, pp. 3361–3368, Singapore, May 2010.

[11] F. Oniga and S. Nedevschi, "Processing dense stereo data using elevation maps: road surface, traffic isle, and obstacle detection," *IEEE Transtractions on Vehicular Technology*, vol. 59, no. 3, pp. 1172–1182, 2010.

[12] Z. Zhang, Y. Wang, J. Brand, and N. Dahnoun, "Real-time obstacle detection based on stereo vision for automotive applications," in *IEEE 5th European DSP Education and Research Conference*, pp. 281–285, Amsterdam, September 2012.

[13] P. H. Batavia and S. Singh, "Obstacle detection using adaptive color segmentation and color stereo homography," in *IEEE International Conference on Robotics and Automation (ICRA)* of *IEEE Computer Society*, pp. 705–710, Los Alamitos, May 2001.

[14] G. Fu, P. Corradi, A. Menciassi, and P. Dario, "An integrated triangulation laser scanner for obstacle detection of miniature mobile robots in indoor environment," *IEEE/ASME Transactions on Mechatronics*, vol. 16, no. 4, pp. 778–783, 2011.

[15] C. Y. Yang, *Intelligent Fall Prevention Shoes Using Camera Based Line-Laser Obstacle Detection System,* [M.S. thesis], National Taiwan University, Taipei, Taiwan, 2014.

[16] T. H. Lin, C. Y. Yang, and W. P. Shih, "Obstacle detection system using line-laser," in *3D Systems and Applications*, Seoul, 2014.

[17] D. A. Winter, A. E. Patla, J. S. Frank, and S. Walt, "Biomechanical walking pattern changes in the fit and healthy elderly," *Physical Therapy*, vol. 70, no. 6, pp. 340–347, 1990.

[18] S. A. Gard, S. C. Miff, and A. D. Kuo, "Comparison of kinematic and kinetic methods for computing the vertical motion of the body center of mass during walking," *Human Movement Science*, vol. 22, no. 6, pp. 597–610, 2004.

[19] P. Fitzpatrick and C. C. Kemp, "Shoes as a platform for vision," in *Proc. of the Seventh IEEE Intl. Symposium on Wearable Computers* of *IEEE Computer Socity*, pp. 231–234, New York, October 2003.

[20] B. R. Frieden, "A new restoring algorithm for the preferential enhancement of edge gradients," *Journal of the Optical Society of America*, vol. 66, no. 3, pp. 44–48, 1976.

[21] R. Harley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Chapter 2, Cambridge University Press, 2003.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 8691505, 12 pages
https://doi.org/10.1155/2017/8691505

Research Article

Atlas-Free Cervical Spinal Cord Segmentation on Midsagittal T2-Weighted Magnetic Resonance Images

Chun-Chih Liao,[1,2] Hsien-Wei Ting,[2,3] and Furen Xiao[1,4]

[1]*Institute of Biomedical Engineering, National Taiwan University, No. 1, Sec. 1, Renai Rd., Taipei City 10051, Taiwan*
[2]*Department of Neurosurgery, Taipei Hospital, Ministry of Health and Welfare, No. 127, Siyuan Rd., New Taipei City 24213, Taiwan*
[3]*Department of Information Management, Yuan Ze University, No. 135, Yuan-Tung Road, Chungli, Taoyuan 32003, Taiwan*
[4]*Department of Neurosurgery, National Taiwan University Hospital, No. 7, Zhongshan S. Rd., Taipei City 10002, Taiwan*

Correspondence should be addressed to Furen Xiao; xfr@dr.com

Received 5 January 2017; Revised 14 February 2017; Accepted 15 February 2017; Published 4 May 2017

Academic Editor: Junfeng Gao

An automatic atlas-free method for segmenting the cervical spinal cord on midsagittal T2-weighted magnetic resonance images (MRI) is presented. Pertinent anatomical knowledge is transformed into constraints employed at different stages of the algorithm. After picking up the midsagittal image, the spinal cord is detected using expectation maximization and dynamic programming (DP). Using DP, the anterior and posterior edges of the spinal canal and the vertebral column are detected. The vertebral bodies and the intervertebral disks are then segmented using region growing. Then, the anterior and posterior edges of the spinal cord are detected using median filtering followed by DP. We applied this method to 79 noncontrast MRI studies over a 3-month period. The spinal cords were detected in all cases, and the vertebral bodies were successfully labeled in 67 (85%) of them. Our algorithm had very good performance. Compared to manual segmentation results, the Jaccard indices ranged from 0.937 to 1, with a mean of 0.980 ± 0.014. The Hausdorff distances between the automatically detected and manually delineated anterior and posterior spinal cord edges were both 1.0 ± 0.5 mm. Used alone or in combination, our method lays a foundation for computer-aided diagnosis of spinal diseases, particularly cervical spondylotic myelopathy.

1. Introduction

The human spinal cord is a long cylindrical structure of the central nervous system extending from the medulla oblongata. Its function is relaying neural signals between the brain and the rest of the body. Residing within the spinal canal formed by the spinal vertebrae, the spinal cord is prone to external compression caused by degeneration, trauma, and so forth. Pathological conditions affecting the spinal cord, also known as myelopathy, lead to motor, sensory, and autonomic dysfunctions, as well as a reduction in quality of life [1]. Among them, cervical spondylotic myelopathy (CSM) is the commonest cause of spinal cord dysfunction in adults globally [2].

Current radiological modality of choice to assess the severity of cervical myelopathy is magnetic resonance imaging (MRI). It provides information about the etiology of canal stenosis, the degree of cord compression, and pathological changes within the cord [3]. Fehlings et al. measured canal compromise on computed tomographic (CT) and T1- and T2-weighted MR images, as well as cord compression on T1- and T2-weighted MR images from patients with spinal cord injury [4]. Based on these methods, experts have developed standardized measurements on midsagittal MR images to quantitatively assess the severity of cord compression in cervical myelopathy in recent years [1–3]. Automation of these measurements requires segmentation of the spinal cord, whether compressed or not, in the original MR images. To our knowledge, no such attempt has been reported.

Current automatic or semiautomatic spinal cord segmentation algorithms focus on multiple sclerosis, which causes atrophy manifested as a decreasing spinal cord area in MR images [5–8]. The earliest semiautomatic method based on an active surface overestimated the cord area in T1-weighted images by approximately 14%, compared to manual outlining [5]. To initialize the algorithm, a human user must mark the approximate cord centerline on a few representative slices. Deformable atlas and Hough transform were employed in newer methods to decrease human intervention used to detect the cord in axial images as well as to improve segmentation accuracy [6, 7]. The Dice coefficients were around 0.9 for the T1-, T2-, and T2∗-weighted images. For spinal cord segmentation in MR images from patients with CSM, these methods may encounter a problem when the cerebrospinal fluid (CSF) spaces outside the cord are compressed secondary to canal stenosis, reducing local tissue contrast.

In the literature, there were only a few atlas-free segmentation methods for the human spinal canal or spinal cord [9, 10]. Archip et al. presented a knowledge-based approach to identify the spinal cord in CT images of the thorax [9]. They constructed a task-oriented anatomical structure map to define the lumbar vertebrae. Although they employed knowledge at incorrect body regions, the results were useable because bony structures are the brightest ones and have fairly stable intensity levels. Kawahara et al. proposed a method to find the globally optimal segmentation of the spinal cord using a high dimensional minimal path search [10]. They represent spinal cord shape principal component analysis. Then, a modified $A*$ minimal path search algorithm in six dimensions was used. Despite dramatically reduced memory requirement, their run-time was between 1 and 5 hours per case.

In this paper, we report an automatic atlas-free algorithm that can perform cervical spinal cord segmentation in standard T2-weighted sagittal MR images without any preprocessing. Human intervention is minimized. Without an atlas, the anatomical knowledge is transformed into constraints employed at different stages of the algorithm. Our method is able to find the spinal cord in images from patients without disruption of the spinal canal. We applied this method to a large number of consecutive patients undergoing a noncontrast MRI study over a 3-month period. The results are presented and evaluated.

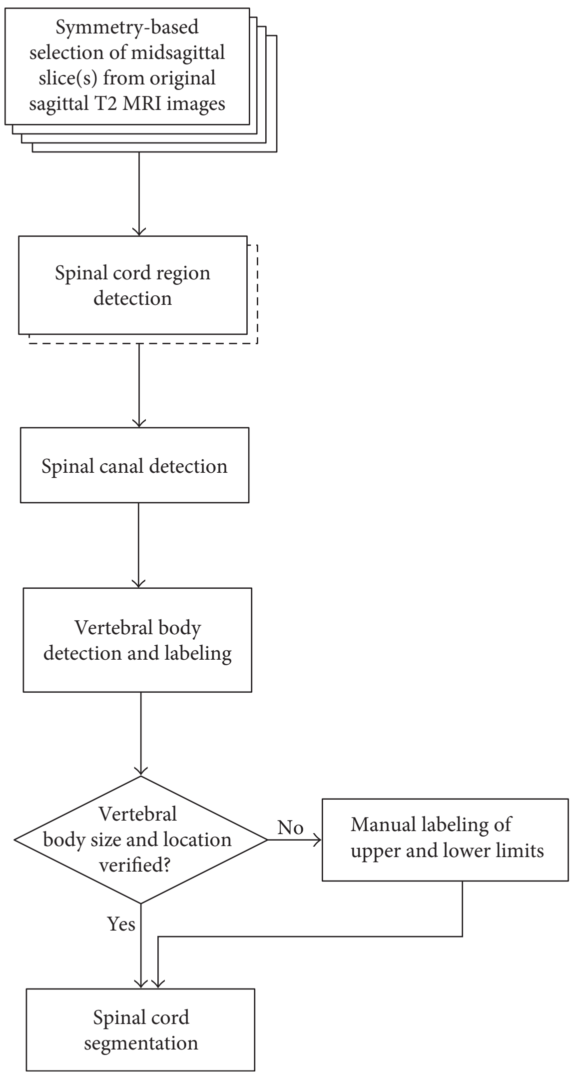


Figure 1: The flowchart of our algorithm.

2. Materials and Methods

2.1. Materials. All adult subjects undergoing noncontrast cervical spine MRI examination from October to December 2015, mainly for CSM, at a regional hospital in Northern Taiwan were retrospectively identified in the database. Patients with a history of cervical spine surgery were excluded. Sagittal T2-weighted images from the subjects were downloaded from the picture archiving and communication system to a personal computer in lossless JPEG format. Our data collection process conformed to the requirements of Institutional Review Board, Taipei Hospital, Department of Health, Taiwan, and was approved as such (TH-IRB-0016-0001).

Image acquisitions were performed on a Siemens Magnetom Avanto 1.5 Tesla MRI scanner (Siemens Healthcare, Erlangen, Germany) using standard coils. Each subject had a T1- and a T2-weighted scan covering the full cervical spinal cord. Parameters for the T2-weighted scan were turbo spin echo sequence, TR = 3300 ms and TE = 95 ms; flip angle = 150°; bandwidth = 223 Hz/voxel; number of averages = 2; and reconstruction diameter = 22 × 22 cm. For sagittal T2 sequence, 3 mm sagittal slices with 0.33 mm gaps between them were planned over the coronal image to cover the whole spinal canal [11]. A saturation band is placed over the anterior, inferior aspect. A total of 13 gray scale images were generated in each sagittal T2 scan. These images are 320 × 320 pixels in size, with a resolution of 0.6875 mm per pixel. The signal intensities (SIs) of the pixels assume a relative scale, stored in 256 gray levels.

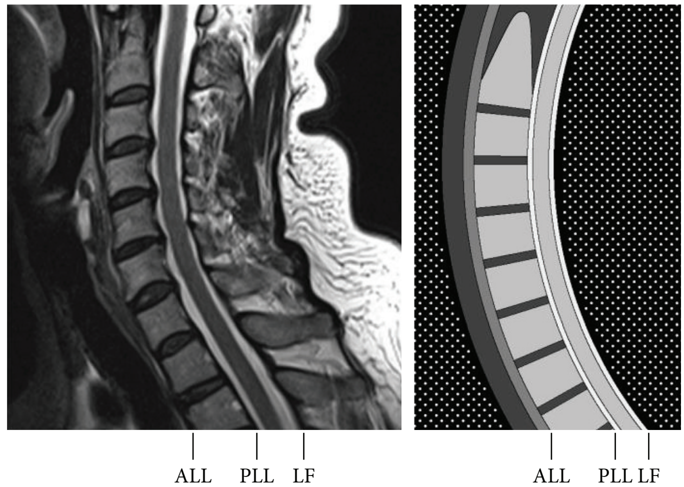


Figure 2: A midsagittal T2-weighted MR image (left) and our schematic drawing showing the spinal cord and its surrounding structures (right). The ligaments, including the anterior longitudinal ligament (ALL), the posterior longitudinal ligament (PLL), and the ligamentum flavum (LF), are deliberately thinned, and the internal architectures of intervertebral disks are neglected.

2.2. Symmetry-Based Selection of the Midsagittal Image. The flowchart of our algorithm is shown in Figure 1. For each MRI data set, we start from selection of one or two midsagittal images based on symmetry between pairs of images. The spinal cord is detected using expectation maximization (EM) and dynamic programming (DP). Only one image is designated as the midsagittal image and undergoes further processing according to the model depicted in Figure 2.

Using DP, the anterior and posterior edges of the spinal canal are detected, as well as the approximate anterior edge of the vertebral bodies (VBs). After thresholding and DP, the VBs and the intervertebral disks were segmented using region growing and then labeled according to their relative sizes. The superior and inferior edges of the cervical VBs were verified by the user and corrected as needed. Finally, the spinal cord is segmented by using DP to detect its anterior and posterior edges.

We define x-, y-, and z-axes as left-right, anterior-posterior (ventral-dorsal), and superior-inferior (cranial-caudal) axes, respectively. Measurements performed on x-, y-, and z-axes are termed width, depth, and height unless otherwise specified. xy, yz, and xz planes correspond to axial, sagittal, and coronal anatomical planes. A sagittal T2-weighted scan contains 13 gray scale images, denoted as I^k, where $k = 1, 2, \ldots, 13$. Let $I^k_{y,z}$ denote the gray level of the pixel at position (y, z) of the kth sagittal image I^k, $0 \le I^k_{y,z} \le 255$ for $1 \le y, z \le 320$.

Similar to the brain, the human vertebral column and the spinal cord are bilaterally symmetric about the intact midsagittal plane (iMSP) [12]. We apply this knowledge to identify the midsagittal MR image, which usually contains the largest anterior-posterior (A-P) diameter of the cervical spinal cord. Let k_{MSP} denote the sagittal position closest to the iMSP. We have to define a difference metric between two images I^p and I^q, denoted as $D(I^p, I^q)$, and then test different trial values of k_{MSP} to find the best one minimizing the global asymmetry quantified using several pairs of corresponding images.

$$k_{\mathrm{MSP}} = \arg\min_k \left(\sum_{j=1}^{13} \frac{D\left(I^j, I^{2k-j}\right)}{m} \right), \tag{1}$$

where m denotes the number of image pairs.

Theoretically, k_{MSP} can assume any real value between 1 and 13 in our images, with an ideal value of 7 for a perfectly positioned patient. However, interpolating images take additional time and generate noisy in-between images, so we only use original images to evaluate global symmetry. Since $D(I^p, I^q)$ is defined only when p and q are integers, k_{MSP} must be an integer or a half-integer. When k_{MSP} is an integer, $I^{k_{\mathrm{MSP}}}$ is the only image near the iMSP. When it is a half-integer, $I^{k_{\mathrm{MSP}}-0.5}$ and $I^{k_{\mathrm{MSP}}+0.5}$ are the two nearest images located on both sides of the iMSP, but they are usually not equally distant from it. To make meaningful comparison using enough number of images, we only evaluate candidate values when there are at least 4 image pairs available; thus, $4 \le k_{\mathrm{MSP}} \le 10$.

Several functions can be chosen as the definition of $D(I^p, I^q)$, including the mean or standard deviation of the SI difference, cross-correlation between corresponding pixels, and the negative of mutual information (MI) using joint histogram. After a pilot study, we found that the standard deviation of the corresponding pixels' gray level differences between I^p and I^q performs best empirically, so $D(I^p, I^q)$ is defined as such. As a result, $D(I^p, I^q) = D(I^q, I^p)$ and $D(I^p, I^p) = 0$ for the same image. After computing $D(I^p, I^q)$ for all 78 image pairs, the k_{MSP} of the given data set can be found.

2.3. Spinal Cord Detection Using Expectation Maximization and Dynamic Programming. Table 1 lists SIs of different tissues on T2-weighted MR images. The spinal cord generally has a smooth contour throughout its course, as

Table 1: Signal intensities of different tissues on T2-weighted MR images. ALL: anterior longitudinal ligament; PLL: posterior longitudinal ligament; LF: ligamentum flavum.

Structure	Component	Signal intensity	Remark
Spinal cord		Isointense	Reference
Cerebrospinal fluid		Hyperintense	
Vertebral body	Cortical bone	Hypointense	May vary
	Bone marrow	Isointense	
	End plate	Hypointense	
Intervertebral disk	Annulus fibrosus	Hypointense	
	Nucleus pulposus	Hyperintense	Decreases with age
Ligaments (ALL, PLL, and LF)		Hypointense	
Air		Very hypointense	

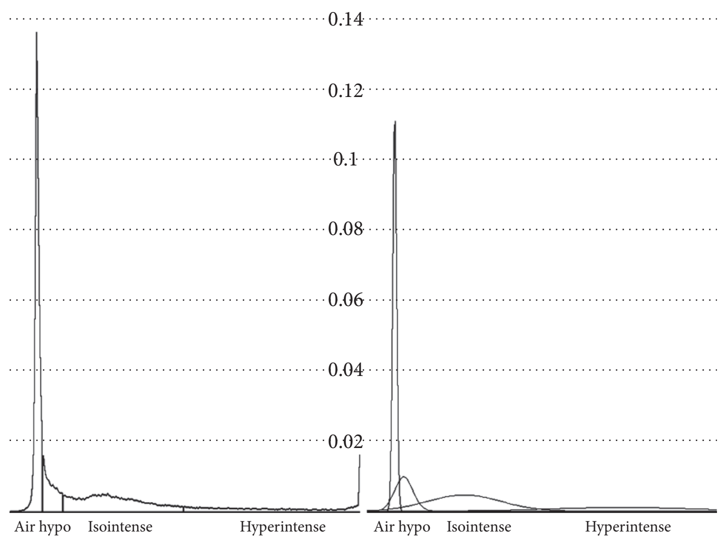


Figure 3: An example of classifying pixels on the histogram (left) after fitting with four Gaussian distributions (right). Horizontal dotted lines denote relative frequency.

shown in Figure 2. It is isointense to the brainstem on all imaging sequences, while the surrounding CSF demonstrates characteristic hyperintensity on T2-weighted images [13]. On images, the spinal cord serves as the reference to classify pixels into hyper-, iso-, and hypointense ones, which have SIs higher than, similar to, and lower than it, respectively.

The normal appearance of the VBs is determined by the ratio of fatty yellow marrow to the hematopoietic red marrow, while their bony cortex demonstrates low SI. The three major components of an intervertebral disk include the nucleus pulposus, annulus fibrosus, and cartilaginous end plate. Only the nucleus pulposus in its core demonstrates high SI due to high water content, which decreases with age. The other two structures demonstrate low SI and are difficult to differentiate from the surrounding vertebral cortex and ligaments.

The ligaments of the spine include the anterior longitudinal ligament (ALL), the posterior longitudinal ligament (PLL), and the posterior ligamentous complex among which the ligamentum flavum (LF) immediately posterior to the CSF-containing dural sac is of our interest. On T2-weighted images, the ALL and the PLL are seen as hypointense bands along the anterior and posterior edges of the vertebral column, while the LF is seen as a hypointense band extending along the posterior edge of the spinal canal (Figure 2).

We employ EM algorithm to cluster the pixels on the given midsagittal MR image according to their gray levels, or SIs. This method is widely used in processing brain MR images [14]. A Gaussian mixture model (GMM) is employed to fit the normalized histogram, as shown in Figure 3. This model assumes that the MR image consists of a number of distinct tissue types from which every pixel has been drawn. The intensities of pixels belonging to each of tissue type conform to a normal distribution, which can be described by a mean, a variance, and the number of pixels belonging to the distribution.

For human experts, visually classifying the pixels into hyper-, iso-, and hypointense ones is enough for making diagnoses. However, we frequently encountered problems in modeling the histogram using only three Gaussians because the hypointense pixels do not assume a perfect Gaussian distribution. Since there is no reason to assume that the hypointense ligament pixels share the same Gaussian distribution as the hypointense air pixels, we used two significantly overlapping Gaussians to fit the hypointense peak, increasing the total number to 4. Because the larger, narrower Gaussian mostly represents air pixels outside the body, we call it "air intense" to represent these very hypointense pixels. The other three Gaussians are called hyper-, iso-, and hypointense, respectively. Using EM, the number of Gaussian distributions is increased sequentially from one to four to fit the normalized histogram. Other details of the algorithm are implemented according to a textbook [15].

In spine images, isointense pixels are the most important as they usually represent the cord. The given gray level is classified isointense if the isointense Gaussian distribution contributes to the largest portion of that part of the histogram, as shown in Figure 3. Other pixels are classified similarly. Those with gray levels lower than the peak of the air intense pixels are automatically classified as such.

Although the EM algorithm always converges, sometimes the "isointense" Gaussian does not accurately represent true isointense pixels, that is, the cord. The SIs of the pixels can be affected, or "modulated," by inhomogeneity of the radiofrequency field, placement of saturation bands, and adjustment by the MR operator [11]. On sagittal T2-weighted images, artifactual longitudinal thin linear hyperintensities are routinely seen, known as Gibbs artifact or truncation artifact [13]. Moreover, quantization of the SI into 256 gray levels also affects the EM process as the hyperintense Gaussian is often truncated if its mean is close to 256. As a result, relying on a fixed threshold to find the isointense pixels inevitably causes problems in some images.

From our pilot studies, we have found that the gray levels of the spinal cord pixels are mostly between 60 and 100. To cope with the errors associated with the EM algorithm, we checked the upper threshold of isointense pixels derived from EM. If it is larger than 127 or smaller than 64, adjustments are made for correct classification. If the threshold is larger than 127, it is recalculated according to the mean and standard deviation of the isointense Gaussian distribution. The resulting value is limited to the range between 127 and 159. On the other hand, if the upper threshold is smaller than 64, indicating three Gaussians assigned to the hypointense and air intense peaks, it is set to 128 without further recomputation.

Dynamic programming (DP) is a method of solving problems by combining the solutions to simpler subproblems [16]. It is typically applied to optimization problems to find a solution with the optimal value. "Programming" in this context refers to the method of tabulating the solutions of the subproblems. When the subproblems overlap, a DP algorithm solves each subproblem just once and then saves its answer in a table, thereby avoiding the work of recomputing it for many times.

In our application, we want to detect some anatomical structures on a midsagittal MR image using DP. We regard this image I as a large 320×320 checkerboard. The fitness or optimality of a given pixel at position (y, z), $f_{y,z}$, can be defined locally using features derived from its gray level, $I_{y,z}$, and from its neighboring pixels. Then, the optimal solution representing the structure to be detected is characterized as the best path B, which is composed of values $\{b_{z_{\text{sup}}}, b_{z_{\text{sup}}+1}, \ldots, b_{z_{\text{inf}}}\}$, representing a series of points $(b_{z_{\text{sup}}}, z_{\text{sup}}), (b_{z_{\text{sup}}+1}, z_{\text{sup}}+1), \ldots, (b_{z_{\text{inf}}}, z_{\text{inf}})$ running from the uppermost row $z = z_{\text{sup}}$ to the lowermost row $z = z_{\text{inf}}$ within the region of interest. The path must be continuous, so one can only move from (y, z) to $(y-1, z+1)$, $(y, z+1)$, or $(y+1, z+1)$.

We define the cumulative fitness values of a given pixel at (y, z), $q_{y,z}$, using $f_{y,z}$ and values from its allowable predecessors, $q_{y-1,z-1}$, $q_{y,z-1}$, and $q_{y+1,z-1}$. For the first row, $q_{y,z_{\text{sup}}}$ is equal to $f_{y,z_{\text{sup}}}$. When the pixel is out of the region of interest bounded by y_{ant} and y_{post}, it is excluded from the best path. In our application, z_{sup} and z_{inf} are constants, while y_{ant} and y_{post} can be functions of z or constants.

$$q_{y,z} = \begin{cases} 0 & \text{if } y < y_{\text{ant}} \text{ or } y > y_{\text{post}} \\ f_{y,z} & \text{if } z = z_{\text{sup}} \\ \max(q_{y-1,z-1}, q_{y,z-1}, q_{y+1,z-1}) + f_{y,z} & \text{otherwise.} \end{cases} \tag{2}$$

Starting from the uppermost row, the table storing of the cumulative fitness values is constructed. An auxiliary table $p_{y,z}$ is also constructed to store the locations of the predecessors of a given point,

$$p_{y,z} = \begin{cases} -1 \text{ if } & q_{y-1,z-1} > q_{y,z-1},\ q_{y-1,z-1} > q_{y+1,z-1} \\ 0 \text{ if } & q_{y,z-1} > q_{y-1,z-1},\ q_{y,z-1} > q_{y+1,z+1} \\ +1 \text{ if } & q_{y+1,z-1} > q_{y-1,z-1},\ q_{y+1,z-1} > q_{y,z-1}. \end{cases} \tag{3}$$

If the maximum occurs in two or more predecessors, $p_{y,z}$ is defined as 0 if $q_{y,z-1}$ is the maximum and as -1 if $q_{y+1,z-1} = q_{y-1,z-1} > q_{y,z-1}$.

To find B, we select the point with the largest cumulative fitness value in the lowermost row and then backtrack its predecessors using the auxiliary table until the first row is reached.

$$b_{z_{\text{inf}}} = \arg\max_{y} \left(q_{y,z_{\text{inf}}}\right), \tag{4}$$

$$b_{z-1} = b_z + p_{b_z,z} \text{ for } z = z_{\text{inf}} - 1, \ldots, z_{\text{sup}}. \tag{5}$$

Let f^w denote the fitness function used to compute the wth best path B^w. In the following sections, we use DP several times to find various anatomical structures represented by $B^1, B^2, \ldots, B^8$ using $f^1, f^2, \ldots, f^8$.

Connected to the brainstem, the spinal cord is the only isointense structure traversing the whole image vertically, as shown in Figure 2. The shape and size of the cord are limited by those of the bony spinal canal, which vary considerably. For patients with cervical spinal canal stenosis, those with a 7.1 mm canal depth, about 10 pixels deep in our images, were more likely to have CSM, whereas patients with a 10.8 mm canal were more likely to be non-myelopathic [17]. Halving this value, we use 5 pixels as a reasonable estimate for the A-P diameter (depth) of the compressed cord.

In a given midsagittal image, the spinal cord can be detected by finding the longest isointense structure of sufficient depth. We define f^1 using classification results of 5 consecutive pixels in y direction,

$$f^1_{y,z} = \sum_{j=-2}^{2} u_{y+j,z}, \quad \text{where } u_{y,z} = \begin{cases} 1 & \text{if } I_{y,z} \text{ is isointense} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The boundaries z_{sup}, z_{inf}, y_{ant}, and y_{post} are set to 1, 320, 1, and, 320, respectively. Then, B^1 can be found using DP.

Since B^1 employs the results of the EM algorithm instead of the original SI, it can be fine-tuned using local SI homogeneity, as defined in f^2. Constants are added to ensure that f^2 has positive values everywhere. Similar adjustments are also used in (8) and (9).

$$f^2_{y,z} = \sum_{j=-3}^{2} \left(65536 - \left(I_{y+j+1,z} - I_{y+j,z}\right)^2\right). \tag{7}$$

Working around B^1, we compare 6 pairs of consecutive pixels to detect the best homogeneous isointense structure B^2, which is likely to be the cord, using DP. z_{sup},z_{inf}, y_{ant}, and y_{post} are set to 1, 320, $B^1 - 40$, and $B^1 + 40$, respectively. Although local zigzagging is common, B^2 is usually closer to the spinal cord along its course than B^1, as shown in Figure 4. Since B^1 and B^2 are not final segmentations, there is no need for them to be "centerlines" of the spinal cord.

Using f^1 and f^2, the exact location of the spinal cord can be detected on the sagittal image. If there is only one midsagittal image selected from the previous stage, the histogram along with pixel classification, B^1 and B^2 are saved and further processing is performed. If there are 2 candidate midsagittal images, we select one whose SI among pixels traversed by B^2 is more stable by comparing their SIs against the moving average derived from 31 neighboring pixels along the path, summed over the lower two-thirds of the image. Setting the range to 31 pixels, we hope to cover one VB and one intervertebral disk to decrease error associated with adjacent structures.

2.4. Ligament Detection Using Dynamic Programming. We define the range where the ligaments, namely ALL, PLL, and LF, may be detected, relative to the location of B^2, using measurements described in the literature [17]. The normal cervical spinal canal has an approximate depth of 15–20 mm, corresponding to 22–30 pixels in our images. The normal lower cervical VBs, including C3, C4, C5, C6, and C7, have approximate depths of 15–20 mm, corresponding to 22–30 pixels. They have approximate heights of 10–15 mm, corresponding to 14–22 pixels. Their normal areas on sagittal images range approximately from 300 to 600 pixels.

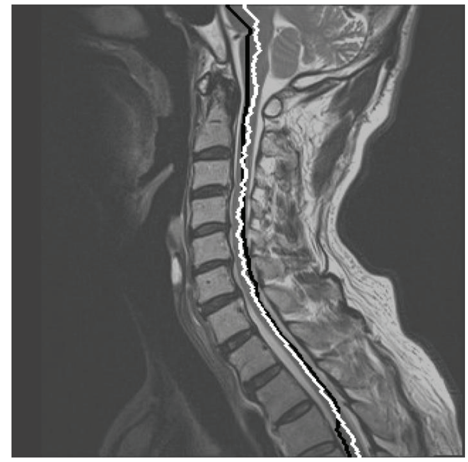


Figure 4: Spinal cord detection using dynamic programming. Near the black line indicating the best path containing the largest number of isointense pixels, the white line denoting the best path traversing the region having the most homogeneous signal intensity is detected.

All ligaments appear hypointense on T2-weighted images (Table 1). Therefore, B^3 and $B^4$are defined in a straightforward fashion.

$$f^3_{y,z} = f^4_{y,z} = \left(256 - I_{y,z}\right)^2. \tag{8}$$

Using DP, B^3 and $B^4$are found anterior and posterior to B^2. They represent PLL and LF, respectively. By setting y_{ant} and y_{post} relative to the location of the detected spinal cord, that is, B^2, we can ensure that no other hypointense structures will become the optimal solution erroneously. Based on normal spinal canal depth, the boundaries for B^3, z_{sup}, z_{inf}, y_{ant}, and y_{post} are set to 1, 320, $B^2 - 30$, and $B^2 - 1$, respectively. Those for B^4 are set to 1, 320, $B^2 + 1$, and $B^2 + 30$.

It is more difficult to find ALL because the air-filled trachea is just anterior to it, separated by the thin isointense esophagus. We define f^5 in a slightly different way. In addition to the hypointense ligament and the cortical bone immediately behind it, we detect 16 isointense bone marrow pixels further posteriorly.

$$f^5_{y,z} = \left(256 - I_{y,z}\right)^2 + \sum_{j=1}^{16} I_{y+j,z}{}^2. \tag{9}$$

Then, the approximate location of ALL, represented by B^5, is found using DP. Based on normal VB depth, the boundaries are set to 1, 320, $B^3 - 60$, and $B^3 - 21$ for B^5. After detecting B^3, B^4, and B^5, we can segment the key regions on the midsagittal image, namely the vertebral column and the spinal canal, as shown in Figure 5.

Although the vertebral column can be reliably detected using DP in most images, additional prevertebral soft tissue regions can be included, which may interfere with separation and detection of individual VBs. For atlas-dependent methods, the VBs and the spinal cord can be detected and labeled after image registration or other template-matching algorithm, usually after manual initialization [18]. Since our method is atlas-free and the only available information at this stage is spinal cord location, we must apply additional techniques to achieve the goal automatically.

2.5. Knowledge-Based Vertebral Body and Intervertebral Disk Detection and Labeling. The vertebral column, with its anterior and posterior edges defined by B^5 and B^3, contains the VBs and the intervertebral disks. The bony cortex of the VB, along with the annulus fibrosus and the end plates of the disks, is hypointense. Other structures of the vertebral column are isointense (Table 1). Therefore, we can construct a histogram of the vertebral column pixels to separate these two groups by finding the corresponding peaks and set the threshold, t_{VB}, at the midpoint. This threshold is usually different from that defined to separate hypointense pixels from isointense ones in the EM process.

To facilitate separation of individual isointense bone marrow regions of the VBs, we need another fitness function f^6 to connect as many hypointense cortical bone and annulus pixels as possible. On the other hand, the posterior half of the VB regions must be retained for region growing algorithm to work.

$$f^6_{y,z} = \left(256 - I_{y,z}\right)^2 + 65536 \sum_{j=1}^{20} v_{y+j,z},$$
$$\text{where } v_{y,z} = \begin{cases} 1 \text{ if } I_{y,z} < t_{\mathrm{VB}} \\ 0 \text{ otherwise.} \end{cases} \qquad (10)$$

The solution of the 6th DP process, B^6, usually lies between B^5 and B^3. It does not correspond to any specific structure but overlaps with B^5 at the anterior edges of the disks. Therefore, we call B^6 "truncated ALL."

After thresholding all pixels between B^6 and B^3 using t_{VB}, we employ exhaustive region growing to segment all isointense regions to find the VBs of the lower cervical spine. For each region, the location, height, depth, and area are calculated. Because some parts of the VB regions are outside the jagged B^6, we consider all regions larger than 150 pixels, that is, larger than half of normal VB size, being valid regions. Then, all regions are sorted according to their z coordinate in preparation for labeling.

The sizes and shapes for the lower cervical and upper thoracic VBs are relatively stable. In contrast, the C1 and C2 vertebrae assume complex shapes. Moreover, they are connected by other ligaments in addition to extensions of the ALL and the PLL, making their detection highly challenging. Despite such complexity, the odontoid process and the body of C2 vertebra usually appear as a connected region collectively, having a total height that averaged 30 mm in adults, about 1.5 times that of the lower cervical VB [19]. We use this knowledge to detect C2.

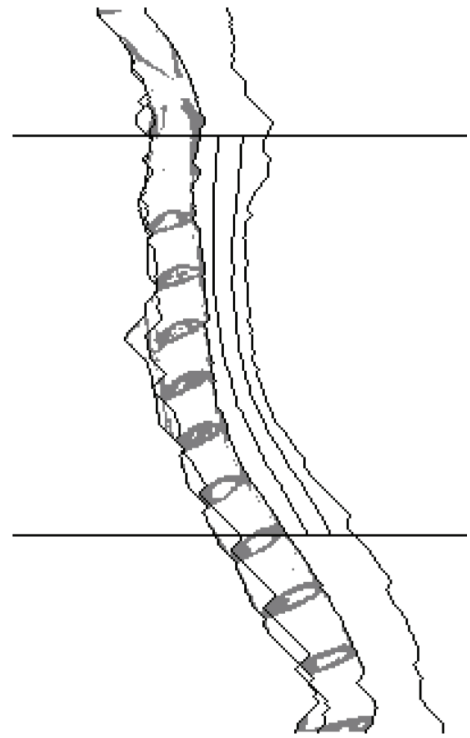


Figure 5: Detecting the edges of different structures using dynamic programming. From left to right: B^5, B^6, B^3, B^7, B^8, and B^4. The horizontal lines denote the superior and inferior edges of the spinal canal. The original MR image is shown in the left part of Figure 7.

After detecting a large region followed by more than 5 valid VB regions inferiorly, the distance between the centers of the first and the second regions on the z-axis is checked. If it is smaller than 50 mm, or 70 pixels, these two regions are considered C2 and C3. Beginning from C3, all "regular" VB regions are labeled. The labeling process continues inferiorly for all VBs detected in the region growing process. The superior and inferior edges of the cervical spinal canal are approximated using the superior edge of the C2 region and the interior edge of the T1 region, respectively. An example is show in Figure 5. To avoid errors associated with vertebral regions at levels above C3 or below C7, we employ extrapolation on the z-axis from centers of the C3 and C7 VBs. The height of C2 is estimated as 1.5 times that of C3, while the height of T1 is estimated as that of C7. We intentionally include the T1 region so that the C7-T1 disk can be detected.

The SIs, sizes, and shapes of intervertebral disks change significantly with aging and various disease processes. Compared to detecting the VBs, detecting the disks is a much more difficult task. Therefore, we only attempt disk detection on images with successfully detected and labeled VBs. In late adulthood, the disks are dehydrated and the hyperintense region is minimal. However, the disk is still a region with heterogeneous SI, prohibiting improvement of segmentation accuracy using simple thresholding.

After removing the VB regions, the "void" regions within the vertebral column should be the disks. We use exhaustive region growing to detect the disk regions regardless of SI. Then, all regions larger than 100 pixels are considered disks and are labeled according to labels of the adjacent VBs. Since our method does not take SIs of the disk pixels into account, it is robust to disk pathologies, which commonly accompany CSM, as illustrated in Figure 6.

Although we have tested and tuned our algorithm in a pilot study, it still failed to identify pertinent structures on some images. Compared to the DP algorithm, the region growing algorithm used to detect VBs and disks is more sensitive to noisy SI. Therefore, we display the VB segmentation and labeling results to the user to allow corrections to be made. For images in which our algorithm fails to find or label the VBs, the superior and inferior edges of the cervical spinal canal can be designated by the user. At this stage, the user can also exclude images in which the complete cervical spinal canal does not exist or is not found by the algorithm from further processing.

2.6. Spinal Cord Border Detection Using a Compound Fitness Function. We perform cord segmentation after specifying the anterior, posterior, superior, and inferior edges of the spinal canal. Since there is no atlas for comparison, we have no information about how the SI of a given pixel is affected by various factors. To alleviate this problem, we apply median filtering to decrease noise [20]. For a given z within the spinal canal, we calculate the median gray level of the spinal cord, c_z, using hypointense and isointense pixels within the range of $B_z^2 - 5$ to $B_z^2 + 5$ in y direction and $z - 15$ to $z + 15$ in z direction.

A compound fitness function is then constructed to maximize the contrast at the edge of the spinal cord, while keeping SIs within the spinal cord region as homogeneous as possible. Four terms representing similarity to the median cord gray level, heterogeneity between adjacent pixels, contrast between cord and noncord pixels, and penalty for passing through noncord pixels are incorporated into f^7,

$$f_{y,z}^7 = f_{y,z}^{7s} + f_{y,z}^{7h} + f_{y,z}^{7c} + f_{y,z}^{7p}. \tag{11}$$

The components are defined as follows. The weighting of each term is defined empirically using another training set of images.

$$\begin{aligned} f_{y,z}^{7s} = 0.5\left(I_{y-1,z} - c_z\right)^2 - 3\left(I_{y,z} - c_z\right)^2 \\ - \left(I_{y+1,z} - c_z\right)^2 - \left(I_{y+2,z} - c_z\right)^2 - \left(I_{y+3,z} - c_z\right)^2, \end{aligned} \tag{12}$$

$$f_{y,z}^{7h} = -\left(I_{y,z} - I_{y+1,z}\right)^2 - \left(I_{y,z} - I_{y+2,z}\right)^2 - \left(I_{y,z} - I_{y+3,z}\right)^2, \tag{13}$$

$$f_{y,z}^{7c} = 0.5\left(I_{y,z} - I_{y-1,z}\right)^2, \tag{14}$$

$$f_{y,z}^{7p} = \begin{cases} -131072, & \text{if point } (y,z) \text{ is air intense or hyperintense} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Figure 6: Vertebral body and intervertebral disk detection and labeling in a midsagittal MR image.

Similarly, f^8 is defined using the same components as f^7, with their constituent pixels in reverse order,

$$f_{y,z}^8 = f_{y,z}^{8s} + f_{y,z}^{8h} + f_{y,z}^{8c} + f_{y,z}^{8p}. \tag{16}$$

Using DP, the anterior and posterior edges of the spinal cord, B^7 and B^8, are detected. The region between them represents the spinal cord on the given T2-weighted midsagittal MR image.

The proposed segmentation method was validated against manual segmentation results. Similar to previous works, we used two measurements to validate the areas and edges resulted from the segmentation process [7]. The Jaccard index was used for quantifying the overlapping between cord regions defined by different observers.

$$J = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{17}$$

where TP, FP, and FN denote the numbers of true-positive, false-positive, and false-negative pixels, respectively. It can be easily converted to Dice coefficient using the relationship $D = 2J/(1 + J)$.

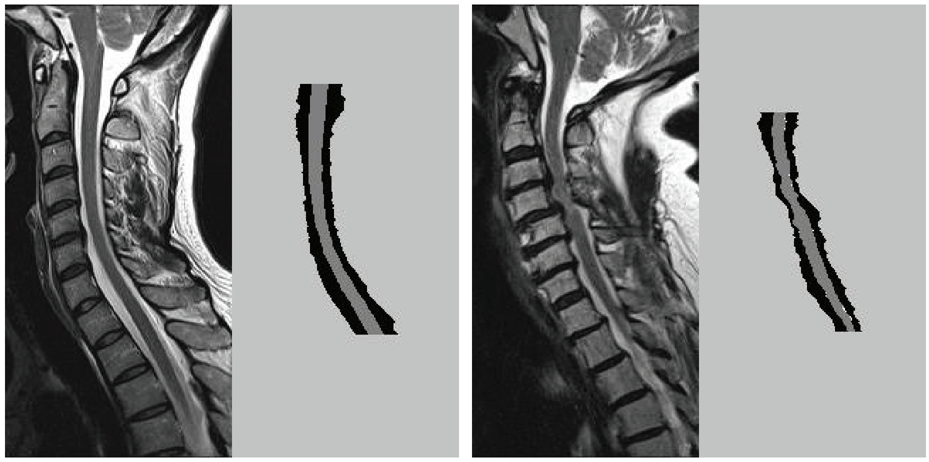


Figure 7: Segmentation results in nearly normal and severely degenerated cervical spines. The left half of each image is the original image containing the cord (gray) and its surrounding cerebrospinal fluid region (white). The right half of each image is the segmentation result. Erroneously classified pixels are shown in white. Components of the epidural space are also shown in some nonstenotic areas, but they are clinically irrelevant.

The Hausdorff distance, which is defined as the maximum distance between two curves, was used to quantify the distance between the anterior and posterior edges of the spinal cord. Three board-certified neurosurgeons performed spinal cord segmentation on the same images used for automatic segmentation. Interobserver agreements between them were calculated. Agreements between each human observer and the automatic method were also calculated. Then, the gold standard was determined using a voting process from three manual segmentations to assess the accuracy of the proposed method.

3. Results

A total of 84 eligible data sets from 84 patients were identified in the hospital database. All 1092 T2-weighted sagittal MR images were downloaded and processed. Our symmetry-based selection algorithm found 156 midsagittal images. These images were reviewed manually and were found to contain the odontoid process, which has an average width of 9 mm and is located near the iMSP [21]. Therefore, all of them were verified as midsagittal images.

Despite successful detection of midsagittal structures, 10 images from 5 data sets were excluded from further processing. In 4 data sets, no single MR image contains the complete cervical spinal canal due to excessive scoliosis, disqualifying the constraints of our algorithm. In one data set, the orientation of the spinal canal was significantly changed due to severe kyphosis related to thoracic wall deformity, prohibiting the DP algorithm to find appropriate solutions. Without the review process, these images would still fail VB detection and labeling and would be rejected for further processing instead of reporting erroneous automatic segmentation results.

From the remaining 79 data sets, a total of 146 images were selected as midsagittal images eligible for spinal cord detection. Among these 79 patients, there were 40 males and 39 females. Their ages ranged from 25 to 85 years, with a mean of 53.5 ± 12.0. After EM and two rounds of DP, the spinal cords were detected in all images. For each data set containing two midsagittal images, the one containing the more homogeneous spinal cord region was retained. After symmetry-based image selection and cord detection, the 7th image was selected in 45 (57%) of 79 data sets. The 6th and the 8th images were selected in 22 (28%) and 10 (13%) patients, respectively. The 5th image was selected in one and the 9th in another.

In 36 (46%) of the 79 images, the original threshold derived from the GMM was suitable for separating isointense pixels from hyperintense ones. For the remaining images, the thresholds were out of range. Automatic threshold adjustment using the isointense Gaussian was done in 43 (54%). In one image, the hyperintense pixels did not form an obvious peak on the histogram, resulting in absence of the pertinent Gaussian, and the upper threshold for isointense pixels was automatically set to 127.

Using DP, the PLLs were detected on all images without problem. The LFs were also detected on all images. After a manual review, small false-positive regions were noted in 6 cases and small false-negative regions in 2. These errors did not affect accuracy of cord segmentation. On the other hand, the results of ALL detection were less stable.

Prevertebral tissues were frequently considered part of the vertebral column.

After truncating the vertebral column region using B^6 and subsequent region growing operations, 67 (85%) images had successful detection and labeling of the VBs. The labels were correct in 66 images. In one image, the congenitally fused C3-4 vertebral bodies were mistaken as C2 and the labels needed to be corrected. Manual designation for superior and inferior edges of the spinal canal was required in this image and in other 12 whose labeling was unsuccessful. Detection of the anterior and posterior edges of the spinal cord within cervical spinal canal was successful in all 79 images. Two examples are shown in Figure 7.

Table 2: Interobserver agreements between the results of spinal cord segmentation by our algorithm and by three human experts, compared using Jaccard indices. Results in mean ± standard deviation and ranges in parentheses.

	Observer 1	Observer 2	Observer 3
Automatic segmentation	0.980 ± 0.015 (0.922~1.000)	0.979 ± 0.015 (0.937~1.000)	0.977 ± 0.015 (0.939~1.000)
Observer 1		0.987 ± 0.010 (0.955~1.000)	0.987 ± 0.010 (0.957~1.000)
Observer 2			0.989 ± 0.010 (0.951~1.000)

Table 3: Interobserver agreements between the results of spinal cord edge detection by our algorithm and by three human experts, compared using Hausdorff distances. Results in mean ± standard deviation and ranges in parentheses.

	Anterior			Posterior		
	Observer 1	Observer 2	Observer 3	Observer 1	Observer 2	Observer 3
Automatic segmentation	1.51 ± 0.74 (0~3.61)	1.51 ± 0.66 (0~3)	1.54 ± 0.68 (0~3.16)	1.50 ± 0.76 (0~4)	1.44 ± 0.78 (0~5.83)	1.58 ± 0.79 (0~5.10)
Observer 1		1.02 ± 0.43 (0~2.24)	1.01 ± 0.38 (0~2)		1.05 ± 0.55 (0~3.61)	1.05 ± 0.44 (0~3)
Observer 2			0.97 ± 0.39 (0~2)			1.01 ± 0.57 (0~3)

The heights of the spinal canal regions ranged from 150 to 231 pixels, with a mean of 187.8 ± 18.8. On average, males have longer canals than females (201.7 versus 173.5 pixels or 139 versus 119 mm) because they are taller. The number of spinal canal pixels ranged from 3000 to 5760, with a mean of 3943 ± 527. They account for only 4% of all pixels in the image. The number of manually segmented spinal cord pixels ranged from 1148 to 2473, with a mean of 1798 ± 250, and the number of automatically segmented spinal cord pixels ranged from 1155 to 2438, with a mean of 1803 ± 251. On average, the area of the spinal cord occupies about 46% of the spinal canal.

The mean gray levels of manually segmented cord pixels on the images ranged from 57.9 to 123.9, with a mean of 76.0 ± 9.3, while the standard deviations of these cord pixels ranged from 9.2 to 21.1, with a mean of 14.1 ± 2. Generally, the SIs of cord pixels decrease significantly as z increases. The mean correlation coefficient between the gray level and z was −0.57 ± 0.21 with a median of −0.63. In 65 of the 79 images, the correlation coefficients were lower than −0.4.

Compared to the gold standard, our algorithm had very good performance. The Jaccard indices ranged from 0.937 to 1, with a mean of 0.980 ± 0.014. Converted to the Dice coefficient, the range was 0.968 to 1, with a mean of 0.990 ± 0.007, better than that described in the previous study [7]. The Hausdorff distances between the automatically detected anterior spinal cord edge and the manually delineated one ranged from 0 to 3 pixels, about 0 to 2 mm, with a mean of 1.44 ± 0.67 pixels. The Hausdorff distances between the automatically detected posterior spinal cord edge and the manually delineated one ranged from 0 to 5.1 pixels, about 0 to 3.5 mm, with a mean of 1.47 ± 0.76 pixels.

Agreements between the results of spinal cord segmentation by our algorithm and by three human experts were shown in Tables 2 and 3. Both interobserver agreement measurements among three human experts were better than those between human and our algorithm.

4. Discussion

We have proposed an algorithm for automatic cervical spinal cord segmentation from original T2-weighted sagittal images. Our method is accurate and robust. All 156 midsagittal images selected from a total of 1092 were confirmed manually. We used EM on the histogram to find the upper and lower thresholds of isointense pixels. Although some adjustments were needed, our algorithm was able to find all spinal cords automatically, whose areas account for only 4% of all pixels in images having complete cervical spinal canals. Similar double threshold-based method was employed in other studies, but cropping regions of interest from the original images was needed before determining the threshold automatically [22].

Our method is completely atlas-free. The anatomical knowledge was built into the algorithm. Despite minimal human intervention, the results of our method were very accurate. The Dice indices and Hausdorff distances were better than those described in the previous studies [7]. In addition, our method is based on sagittal MR slice. These characteristics make our method complementary to current atlas-dependent methods based on axial images. Clinically useful metrics including cord compression and canal compromise described in [3] can be derived automatically on midsagittal MR images. Although similar knowledge of the spinal cord was incorporated in the methodology described previously [10], our emphasis other than anatomical structures, as detailed in Figure 2, has not been proposed.

When finding the MSP slices, we empirically define the difference metric of image pairs using the standard deviation of the corresponding pixels' gray level differences. However, the most commonly used tool for measuring image similarity is MI [14, 23]. We consider the slightly inferior performance of MI related to uncorrected radiofrequency field inhomogeneity and other artifacts as described in Section 2.3.

The stability of the EM algorithm is lower than we had expected. In addition to aforementioned sources of errors that also destabilize MI, adjustment of the histogram, or "windowing," which helps radiologist reading the images, may also affect EM. After windowing, many hyperintense pixels stuck at the highest gray levels and their gray level distribution is no longer Gaussian. The adjusted threshold was at the allowed maximum in 6 of 43 images whose upper thresholds for isointense pixels were adjusted. In other images, the hyperintense pixels were also too heterogeneous to allow EM to fit a stable Gaussian for them. As a result, the adjusted threshold was at the allowed minimum in 32 images.

We use dynamic programming as the main tool used for detecting anatomical structures and their edges. When the cumulative fitness values $q_{y,z_{\inf}}$ are the same, there may be more than one optimal solution. Compared to the detection of PLL and the LF, detecting the ALL appeared much more difficult using DP. Variations of the prevertebral anatomy may play a role. Incorporating such knowledge may improve the segmentation accuracy for VBs and disks. It takes less than one minute to find the midsagittal slice and only seconds to segment the spinal cord because only two dimensions were used in the searching process. Although not directly comparable, the speed of our clinically oriented one-slice algorithm is considerably faster than the previous one using six dimensions [10].

We constructed two compound functions, f^7 and f^8, to detect the edges of the spinal cord. After considering various aspects affecting tissue contrast between the cord and its surrounding tissues, they seemed rather robust. Within the normal spinal canal, these functions accurately detect the interface between the isointense spinal cord and the hyperintense CSF region. When the CSF space disappears as compressed by the severely stenotic canal, the same functions can detect the interface between the isointense cord and the hypointense ligaments, as illustrated in Figure 7. However, when the width of the CSF region is minimized to one pixel, the partial volume effect may render it isointense, resulting in false-positive results.

Based on our algorithm, used alone or combined with others, one can develop a computer-aided diagnosis system capable of massive screening on cervical spine diseases, particularly CSM. During the review of the automatic spinal cord segmentation results, the human observers also evaluated the severity of canal stenosis. In most patients with moderate and severe stenosis, the changes in the anteroposterior diameter of the spinal cord are limited. On the other hand, changes in sizes of the CSF spaces are much more striking. If the cord diameter is used as the sole parameter measured in patients with CSM, disease severity may be underestimated. Therefore, some experts advocate correlating routine MR images to flexion-extension MR images if the diagnosis is in doubt [24].

There are several limitations to our algorithm that deserve mention. On the given midsagittal MR image, we used the spinal cord as the very first feature to be recognized and processed. Therefore, the algorithm cannot be applied to lower lumbar spinal levels, where the cauda equina composed of multiple nerves is only a neural structure within the canal. Although small regions of SI change within the spinal cord region caused by CSM did not affect its accuracy, our algorithm can fail when there are large cord lesions spanning long spinal levels. For the DP algorithm to detect the anatomical structures, a continuous spinal canal with the structures being aligned roughly and vertically is required. If the canal is disrupted by trauma, tumor, or other pathologies, modifications of our algorithm are required for it to work properly. Parameters of our algorithm must be tuned before application to other anatomical regions, such as thoracic and upper lumbar spines, as well as before application to other MRI scanners.

5. Conclusion

Automatic segmentation of the spinal cord and CSF in MR images remains a difficult task. We have presented an automatic method of spinal cord segmentation on sagittal T2-weighted images employing EM, DP, and region growing algorithms. Relevant anatomical knowledge is transformed into constraints in the algorithm enabling it to being atlas-free. Our method is accurate and robust and requires minimal human intervention. Used alone or combined with other methods, it lays foundation for computer-aided diagnosis of spinal diseases, particularly degenerative ones.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research reported in this publication was supported by the National Taiwan University Hospital, Grant UN106-059.

References

[1] A. Nouri, L. Tetreault, J. J. Zamorano et al., "Role of magnetic resonance imaging in predicting surgical outcome in patients with cervical spondylotic myelopathy," *Spine*, vol. 40, no. 3, pp. 171–178, 2015.

[2] L. Tetreault, C. L. Goldstein, P. Arnold et al., "Degenerative cervical myelopathy: a spectrum of related disorders affecting the aging spine," *Neurosurgery*, vol. 77, Supplement 1, pp. S51–S67, 2015.

[3] A. Karpova, R. Arun, A. M. Davis et al., "Reliability of quantitative magnetic resonance imaging methods in the assessment

of spinal canal stenosis and cord compression in cervical myelopathy," *Spine*, vol. 38, no. 3, pp. 245–252, 2013.

[4] M. G. Fehlings, S. C. Rao, C. H. Tator et al., "The optimal radiologic method for assessing spinal canal compromise and cord compression in patients with cervical spinal cord injury Part II: results of a multicenter study," *Spine*, vol. 24, no. 6, pp. 605–613, 1999.

[5] M. A. Horsfield, S. Sala, M. Neema et al., "Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: application in multiple sclerosis," *NeuroImage*, vol. 50, no. 2, pp. 446–455, 2010.

[6] M. Chen, A. Carass, J. Oh et al., "Automatic magnetic resonance spinal cord segmentation with topology constraints for variable fields of view," *NeuroImage*, vol. 83, pp. 1051–1062, 2013.

[7] B. De Leener, S. Kadoury, and J. Cohen-Adad, "Robust, accurate and fast automatic segmentation of the spinal cord," *NeuroImage*, vol. 98, pp. 528–536, 2014.

[8] B. De Leener, M. Taso, J. Cohen-Adad, and V. Callot, "Segmentation of the human spinal cord," *Magma*, vol. 29, no. 2, pp. 125–153, 2016.

[9] N. Archip, P. J. Erard, M. Egmont-Petersen, J. M. Haefliger, and J. F. Germond, "A knowledge-based approach to automatic detection of the spinal cord in CT images," *IEEE Transactions on Medical Imaging*, vol. 21, no. 12, pp. 1504–1516, 2002.

[10] J. Kawahara, C. McIntosh, R. Tam, and G. Hamarneh, "Globally optimal spinal cord segmentation using a minimal path in high dimensions," in *IEEE 10th International Symposium on Biomedical Imaging (ISBI)*, San Francisco, CA, USA, 2013.

[11] M. Elmaoglu and A. Celik, *MRI Handbook: MR Physics, Patient Positioning, and Protocols*, Springer, New York, NY, 2012.

[12] H. Gray and C. Clemente, *Gray's Anatomy of the Human Body*, Lippincott Williams & Wilkins, Philadelphia, PA, 1984.

[13] M. F. Reiser, W. Semmler, and H. Hricak, *Magnetic Resonance Tomography*, Springer, Berlin, 2008.

[14] R. S. J. Frackowiak, K. J. Friston, C. D. Frith et al., *Human Brain Function*, Academic Press, San Diego, CA, 2004.

[15] D. A. Forsyth and J. Ponce, *Computer Vision: a Modern Approach*, Pearson Education, Upper Saddle River, NJ, 2003.

[16] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 3rd edition, 2009.

[17] E. C. Benzel, *Spine Surgery: Techniques, Complication Avoidance, and Management*, Elsevier/Saunders, Philadelphia, PA, 2012.

[18] E. Ullmann, J. F. Pelletier Paquette, W. E. Thong, and J. Cohen-Adad, "Automatic labeling of vertebral levels using a robust template-based approach," *International Journal of Biomedical Imaging*, vol. 2014, Article ID 719520, p. 9, 2014.

[19] C. Cokluk, K. Aydin, C. Rakunt, O. Iyigun, and A. Onder, "The borders of the odontoid process of C2 in adults and in children including the estimation of odontoid/body ratio," *European Spine Journal*, vol. 15, no. 3, pp. 278–282, 2006.

[20] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, Boston, MA, 1992.

[21] M. Singla, P. Goel, M. S. Ansari, K. S. Ravi, and S. Khare, "Morphometric analysis of axis and its clinical significance-an anatomical study of Indian human axis vertebrae," *Journal of Clinical and Diagnostic Research*, vol. 9, no. 5, pp. AC04–AC09, 2015.

[22] M. M. El Mendili, R. Chen, B. Tiret et al., "Fast and accurate semi-automated segmentation method of spinal cord MR images at 3T applied to the construction of a cervical spinal cord template," *PLoS One*, vol. 10, no. 3, article e0122224, 2015.

[23] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[24] D. Zeitoun, F. El Hajj, E. Sariali, Y. Catonne, and H. Pascal-Moussellard, "Evaluation of spinal cord compression and hyperintense intramedullary lesions on T2-weighted sequences in patients with cervical spondylotic myelopathy using flexion-extension MRI protocol," *The Spine Journal*, vol. 15, no. 4, pp. 668–674, 2015.

Hindawi
Journal of Healthcare Engineering
Volume 2017, Article ID 7460168, 13 pages
https://doi.org/10.1155/2017/7460168

Research Article

A Fast SVM-Based Tongue's Colour Classification Aided by *k*-Means Clustering Identifiers and Colour Attributes as Computer-Assisted Tool for Tongue Diagnosis

Nur Diyana Kamarudin,[1,2] Chia Yee Ooi,[1] Tadaaki Kawanabe,[2] Hiroshi Odaguchi,[2] and Fuminori Kobayashi[1]

[1]*Embedded System Research Laboratory, Department of Electronics System Engineering, Malaysia-Japan International Institute of Technology, Kuala Lumpur, Malaysia*
[2]*Oriental Medicine Research Center, Kitasato University, Minato, Japan*

Correspondence should be addressed to Nur Diyana Kamarudin; east.diyana@gmail.com

Received 2 December 2016; Revised 8 February 2017; Accepted 8 March 2017; Published 20 April 2017

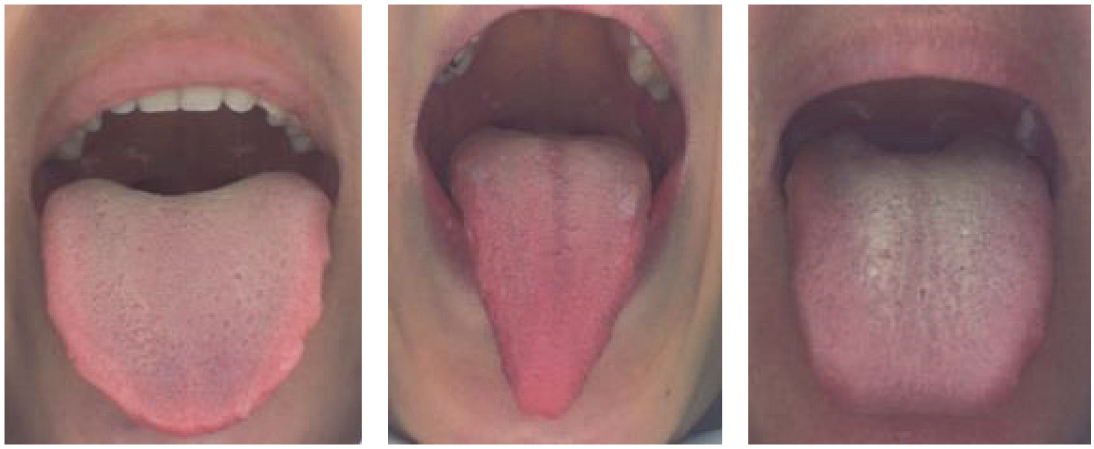
Academic Editor: Junfeng Gao

In tongue diagnosis, colour information of tongue body has kept valuable information regarding the state of disease and its correlation with the internal organs. Qualitatively, practitioners may have difficulty in their judgement due to the instable lighting condition and naked eye's ability to capture the exact colour distribution on the tongue especially the tongue with multicolour substance. To overcome this ambiguity, this paper presents a two-stage tongue's multicolour classification based on a support vector machine (SVM) whose support vectors are reduced by our proposed *k*-means clustering identifiers and red colour range for precise tongue colour diagnosis. In the first stage, *k*-means clustering is used to cluster a tongue image into four clusters of image background (black), deep red region, red/light red region, and transitional region. In the second-stage classification, red/light red tongue images are further classified into red tongue or light red tongue based on the red colour range derived in our work. Overall, true rate classification accuracy of the proposed two-stage classification to diagnose red, light red, and deep red tongue colours is 94%. The number of support vectors in SVM is improved by 41.2%, and the execution time for one image is recorded as 48 seconds.

1. Introduction

Tongue diagnosis is said to be the most active research in the advancement of complementary medicine compared to other diagnosis fields such as pulse and abdominal palpation [1]. For several decades, the objectification of tongue's colour, texture, and geometry analysis as well as pathological feature disease correlations have been researched thoroughly to achieve standardization in clinical practice and to improve the existing technology of computerized tongue image analysis device. In traditional east-Asian medicine, precise tongue colour quantification is essential to predict patients' illness caused by physical and mental disharmony. Hence, tongue colour provides beneficial information on blood congestion, water imbalance, and psychological problems [2]. In *Kampo medicine* as well as traditional Chinese medicine (TCM), tongue colour is mainly classified into three colors such as red, light red, and deep red as illustrated in Figure 1.

In 2010 and 2013, a tongue colour gamut descriptor has been proposed by several researchers using one class SVM [3, 4]. This proposed work suggested that the tongue colour gamut is very narrow and comprises of different types of identical colour; thus, there are many overlapping and similar pixel values that exist. By using the naked eye, the colour information on different regions of a tongue body (or substance) might look almost similar. Nevertheless, when we clustered the tongue body image into distinguished pixels using clustering algorithm, there exist several clustered regions with different colour information on a tongue. These clustered colour information is very useful in order



(a) (b) (c)

Figure 1: Tongue colour samples: (a) light red, (b) red, and (c) deep red.

to determine what are the most contributing colour or area on a tongue body for tongue colour evaluation and diagnosis. However, in order to choose the most informative clustered pixels to be used as input features in a classifier, an effective feature selection method is necessary. According to [5], there are several problems during preprocessing and data selection that can reduce the performance of a classifier:

(i) Impossible features or values have been inputted as training examples.

(ii) Several values have not been inputted during training (missing values).

(iii) Redundant or irrelevant data have been included as training examples.

Similar researches in [6–9] have been discussed to sift out redundant information from the samples used in the classifier during data selection. Nevertheless, some of these works did not consider the basis theory of the sample selection and the possibility of misclassification or misclustering during sample selection. Moreover, too many features included in the training examples have led to complex descriptive feature mapping in the classifier [10]. Currently, there are many researches in computerized tongue image analysis system that utilized machine learning techniques aiming to have better accuracy rate and lesser runtime. Yet, most of the journals or research works have been reported on a trade-off between these two parameters (computation time and accuracy). Several works only reported on the optimization of classification accuracy but not the execution time [10–13].

Therefore, creating a classifier by taking into account the most informative or meaningful features to improve the response time and generalization ability is crucial. This paper presents a two-stage classification system for tongue colour diagnosis aided with the devised clustering identifiers and proposed colour range that can improve the classification accuracy and classifier's response time. Data selection method in our case (*k*-means clustering) is based on tongue colour pattern and area using traditional medicine perspective. The purpose of these clustering identifiers is to collect only meaningful tongue features to reduce the classifier runtime via the reduction of support vectors and outliers. Besides, red colour range in the second stage of classification reduces the possibility of overlapping pixels in the new examples that can boost up the classification accuracy. To the best of our knowledge, this is the latest tongue colour diagnosis system that promotes high accuracy and fast response time classification system which considered the most contributing colour and area analysis using *k*-means, SVM, and colour range. This novel two-stage classification system outperformed the conventional SVM by 20% in terms of computational time and 15% in terms of classification accuracy.

2. Previous Works

Feature extraction or data transformation is a process of transforming a raw feature data into quantitative data structure or patterns for training accessibility [14]. Feature detection and extraction is an essential procedure in tongue diagnostic system after segmentation procedure as some of the tongue features accumulated beneficial information related to the internal body system and health. During tongue manifestation, the medical practitioners examined the tongue's features such as tongue coating, teeth marks, prickles, purple spot and blue spot, or any abnormal features on the tongue to predict the condition of their patient. For instances, teeth marks on a tongue can be related to the dehydration of fluid in the body, substance with existence of purple spot can be related to the blood congestion inside the body, too many obvious prickles are related to the appendicitis [15], and more feature disease or disharmony state of the body correlations can be predicted. Hence, an accurate feature detection or extraction algorithm is crucial to examine the image in the feature's region comprehensively to acquire useful information in terms of its colour or texture.

To establish the standard of tongue diagnosis, the standardization fundamental in defining the most suitable informative features for any significant diseases is also essential. There are several reported works using hyperspectral images to attain meaningful tongue features for further analysis such as coating colour and sublingual veins [16–18]. Even though the hyperspectral image sensing is insensitive to various illumination and albedo effects, the high cost of the hyperspectral camera is a limitation. Since decades, the objectification of tongue colour, texture, and geometry analysis as well as pathological feature disease correlations have been researched comprehensively to establish a computerized tongue diagnosis device. Moreover, to determine the most important features that can exclusively contribute to improve performance of the classifier is very crucial. Some related works are reported in [6–9] using hybrid classifiers to sift out several redundant features and adopted only the most meaningful instances to boost up the performance of the classifier in terms of its response time and classification accuracy. Through this motivation, this technique will reduce the number of training examples and lead to the number of support vectors reduction, thus speeding up the classifier's response time. However, some of these works have not considered the possibility of misclassification or misclustering during sample selection. In addition, the number of clusters produced by k-means clustering is always the same as the number of classes produced by SVM; hence, it cannot be applied in certain cases where SVM classes are less than the clusters of k-means which may occur in some classification problems. To the best of our knowledge, none of them have been implemented on image applications.

Because there is abundance of tongue features or pathological details that have been accumulated via thousands of images progressively, the innovation in decision support system and intelligent image analysis has evolved for accurate and fast classification and diagnosis. In general, there are three learning algorithms which have been implemented using machine learning such as supervised, unsupervised, and semisupervised algorithms. By utilizing these algorithms, an accurate classification system to classify the most informative features with high generalization ability is desired. There are several works reported in feature disease classifications aiming to predict the mapping relationships between tongue features and diseases [10, 11, 15, 19–21]. There are several researches that utilized neural network in categorizing the tongue features [18, 22–24]. Nonetheless, the two most applied classifiers in tongue diagnosis field are Bayesian network classifier [10, 25] and SVM-based classifier [3, 11–13, 26, 27]. Even though there are sufficient training examples from the textural and chromatic properties of a tongue used in the classifier, the accuracy in some reported works needs to be improved [10, 28].

In [26], an SVM-based algorithm called transductive support vector machine (TSVM) is proposed by combining the labelled and unlabelled samples of tongue features as training examples to reduce the human labour and improve the classifier's accuracy since the unlabelled samples help to provide much more classification information during the training process. Nevertheless, there are several classification problems of ambiguous separating boundaries between the classes. This is because there is no model selection method has been made prior to the classification during the training process. The more unlabelled samples are included, the noisy the data will be. Hence, the study on the selection method of unlabelled samples remained as a future research. To compare the performances among classifiers, this research paper has investigated five types of machine learning algorithms to foresee the performance of these algorithms in terms of tongue's features and classification [11]. Five different machine learning algorithms including ID3 (based on decision tree), J48 (based on decision tree), naive Bayes (based on Bayesian network), BayesNet (based on Bayesian network), and sequential minimal optimization (based on SVM) were applied to a tongue dataset of 457 instances. Their comparison results have shown that the SVM-based algorithm has relatively the best performance. However, with the abundance of accumulated tongue features in the near future, the limitations of conventional SVM algorithm are on its speed and size. For a similar generalization performance, SVM response time is slower than other neural network algorithms [29]. The computational complexity that is linear with the number of support vectors is an unsolved problem [30]. To date, the issue on how to choose a good kernel functions in a data-dependent way [31] and how to control the selection of support vectors has been researched thoroughly especially in a noisy and continuous data [32].

3. Materials and Method

3.1. Clinical Data Collection. There are a total of 300 tongue images after coating eliminations have been accumulated during clinical practice in the Oriental Medicine Research Centre, Kitasato University in Japan. All tongue images in this proposed research were taken by tongue image analyzing system (TIAS) that was invented by the Chiba University, Japan. TIAS is a closed box acquisition system that is used to capture the tongue image under stable condition in terms of illumination condition and tongue's position. There are several components implemented in TIAS such as

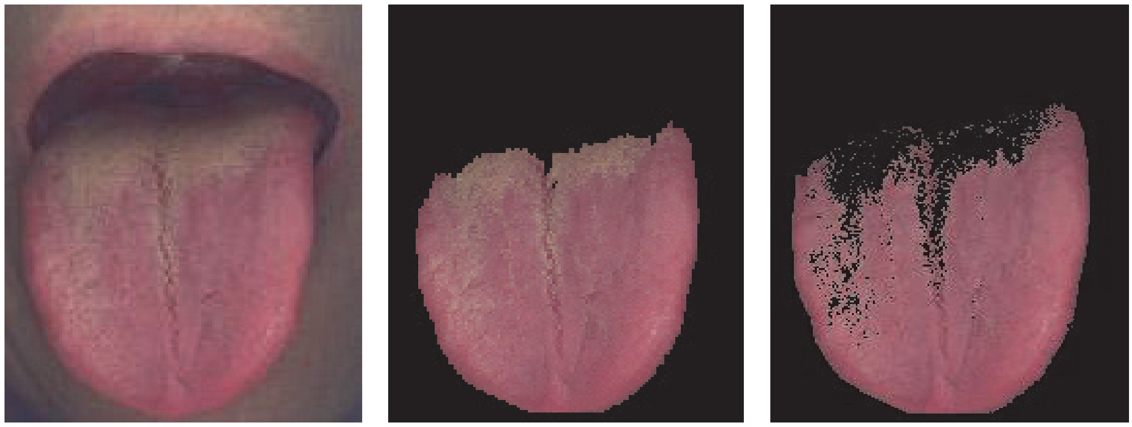
(i) halogen lamps as illuminators with high colour temperature to acquire adequate tongue colour information,

(ii) integrating sphere which is a hollow cavity shaped with coated interior to produce equal distribution of light rays on a tongue,

(iii) 1280×1024 pixels high-speed charged couple digital (CCD) camera to capture high-resolution 24 bit RGB (redness, greenness, and blueness) tongue images,

(iv) 24-colour chart for colour correction purposes.

All the images implemented underwent colour correction procedure to maintain high colour reproducibility outcomes. Around 300 tongue images after coating elimination in [33] are used in k-means clustering procedures and 600 features or instances of clustered results or what we called clustering identifiers are used as training examples in SVM.



(a) (b) (c)

Figure 2: (a) Raw image, (b) image after segmentation, and (c) image after coating removal.

3.2. Fast SVM Aided by k-Means Clustering Identifiers and Colour Attributes. This section describes our full procedures on our proposal of two-stage classification system by implementing an SVM classifier aided with clustering identifiers and red colour range. The preclassification starts with the most contributing colour and area analysis using k-means clustering to develop clustering identifiers that can be used as training examples to recognize tongue colour in SVM. This automatic classification system is proposed to classify tongue images into light red tongue, red tongue, and deep red tongue. In the first-stage classification, tongue images after segmentation and coating removal such as in Figure 2 are fed to k-means clustering algorithm with $k = 4$ to determine the most contributing colour and area on the tongue image.

The outcome of the clustering procedure is the cluster image that is divided into background, red/light red, deep red, and region with transitional pixel clusters. The two most informative clusters or clustering identifiers are red/light red and deep red clusters. These two clustering identifiers will be used as training examples in SVM to preclassify the deep red tongue or red/light tongue. In the second-stage classification or final classification, the red/light red tongue images are further classified into red tongue and light red tongue based on the red colour range derived from our databases. The detail on how we choose the most informative clusters will be discussed in the next section.

3.2.1. The Most Contributing Colour and Area Analysis Using k-Means Clustering Algorithm. The most important step in our proposed classification technique is a sampling strategy of bag-of-features reduction (or also known as feature selection) by applying clustering algorithm to define the most contributing colour and area on tongue images. Feature selection is a procedure of selecting an informative subset of nonredundant features among the original or transformed ones usually for efficiency purposes [34]. This technique is mainly proposed to define the most meaningful colour regions or area (or clusters) that can contribute to diagnose the tongue colour (red, light red, and deep red). Moreover, it is also can be considered as a down sampling technique to eliminate the less contributing colour so that the computational complexity is reduced. Clustering is the process of partitioning a group of data points into a small number of groups. The goal is to assign a data label to each data point and separate the data points to every cluster that depends on the similar label. Given a set of n data points in d-dimensional space R^d and an integer k, k-means clustering determines a set of k centre points in R^d such that the mean squared distance from each data point to its nearest centre point is minimized. In image processing analysis, each pixel in the input image was treated as an object that has a location in space. k-means clustering is usually implemented to solve the image segmentation or feature extraction and feature classification. In our proposed work, k-means clustering algorithm has been implemented with *Lab* colour space to produce informative cluster information and to identify some hidden patterns that can be used as training pixels in SVM to diagnose the tongue colour.

In east-Asian medicine perspective, the tongue accumulates several valuable information regarding the properties, location, and development and prognosis of a disease [35]. The tongue is usually diagnose by observing its bilateral edge regions [2]. Nevertheless, according to TCM, there are five regions of the tongue that can be useful to relate directly or indirectly to the internal organs such as bilateral edges (related to liver and gallbladder), tip (heart and lung), centre (spleen and stomach), and root (kidney). To this motivation, we want to determine the most contributing colour and area on a tongue body that can be used to diagnose the tongue colour accurately. Apart from tongue's root regions, we have discovered that by using clustering algorithms, the tongue regions are distinguished via the edges, the tip of the tongue, and the centre of the tongue. Moreover, we have observed that most of the pixels are clustered at the tip, bilateral edge, and the whole tongue's region (tip, edges, and centre). Hence, we have adopted the two most informative clusters or also known as clustering identifiers which represent the pixel distribution around the tongue's tip and edges which we called the highest colour distance from black pixel $(0, 0, 0)$

and the highest number of pixel cluster which has been distributed sparsely across the tip, edges, and centre of a tongue. We named them as maximum colour distance identifier and maximum pixels' coverage area identifiers to identify the red/light red cluster and deep red cluster, respectively. This procedure was done along with the verification of several Oriental medicine practitioners at Kitasato University, Japan.

The k-means clustering formula is described using cluster centroid detection given by

$$\Delta_{\text{distance metric}} = \sqrt{(L_1{}^* - C_1{}^*)^2 + (a_1{}^* - C_2{}^*)^2 + (b_1{}^* - C_3{}^*)^2}, \tag{1}$$

where $(C_1{}^*, C_2{}^*, C_3{}^*)$ is the centroid of each cluster.

We have tested several possible number of clusters and observed the outcomes with the practitioner's recommendation. By using $k = 3$, the mixture of red, light red, and deep red pixels on one cluster was detected by observing the centroid value of each cluster and the pixels are vaguely distinguished. Whilst, by using $k = 5$, the redundant clusters sharing similar average colour were detected or, in other words, the colour distance between each cluster is not significant. After several experiments, we have chosen $k = 4$ because tongue substance usually comprises of three main colours which is red, light red, and deep red. The other one cluster left is reserved for a background. Moreover, by looking at the perspectives of pixel's distribution around four tongue regions (left and right edges, tip, and the centre) mentioned previously, we found that $k = 4$ is the most accurate cluster number that characterised the distribution. The practitioners confirmed all clustered images on an IPS (in-plane switching) monitor (ColourEdge CG246, EIZO®). The development of clustering identifiers can help in boosting up the performance of SVM via the reduction of noisy pixels and outliers in the training data.

3.2.2. k-Means Clustering Identifiers. Based on our observation during the process of determining the number of cluster, k using resulting clusters of k-means, we have recognized two hidden patterns or we called it as identifiers to identify the clusters that contained deep red region and red/light red region. These identifiers are devised based on (i) maximum colour distance from black pixel $(0, 0, 0)$ and (ii) maximum pixel's coverage area. Moreover, based on the prelabelled images beforehand, the output clusters from the k-means were also measured in terms of their average colour value in *Lab* colour space to develop the red colour range identifier to be used in the second-stage diagnosis.

Maximum colour distances from black pixel or chromatic intensity are determined based on (2) as

$$\Delta_{\text{colour distance}} = \sqrt{(L_2{}^* - L_1{}^*)^2 + (a_2{}^* - a_1{}^*)^2 + (b_2{}^* - b_1{}^*)^2}, \tag{2}$$

where $(L_1{}^*, a_1{}^*, b_1{}^*)$ is $(0, 0, 0)$. Using this identifier, pixels with more red and less blue (higher $a^*$and b^*) concentration (red and light red) can be easily detected compared to pixels with less red and more blue (lower $a^*$and b^*) concentration because these pixels have longer distance from the black pixel. In our *Lab* colour analysis on substance colour after coating removal, red and light red tongue has been observed to have a domination of a^*, b^* (chromatic) pixels compared to deep red tongue that was influenced by black colour as their distance to black pixel is nearer [33]. Hence, after series of training procedures in SVM, this identifier has the best accuracy to classify red/light red tongue.

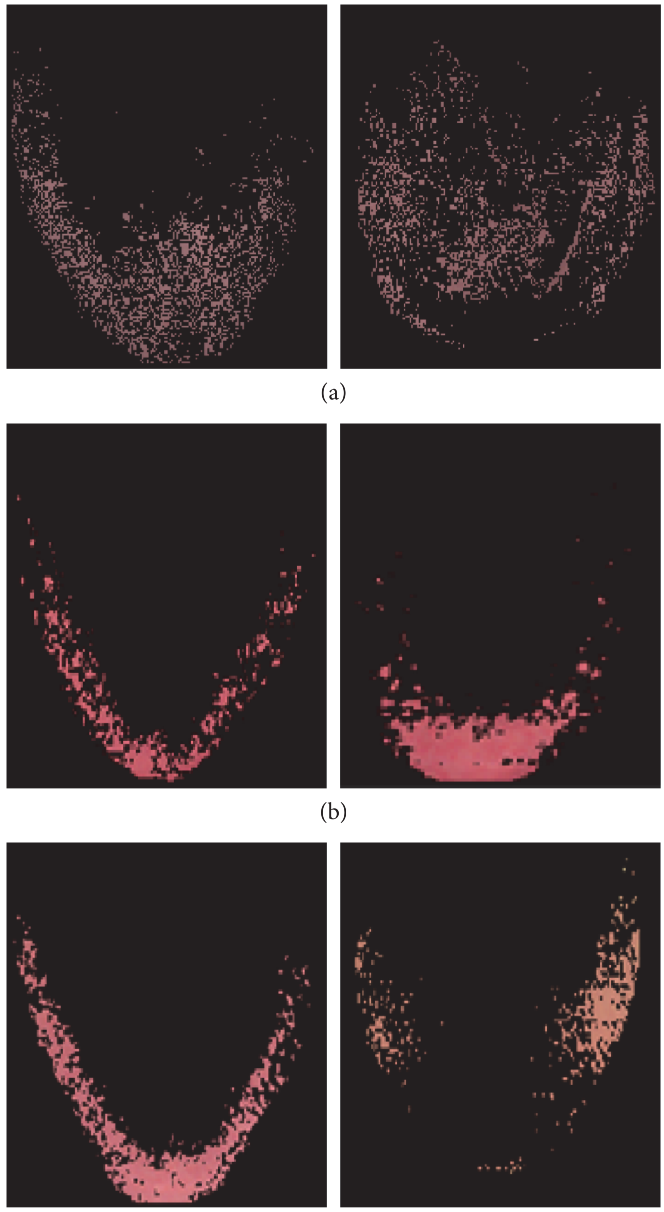
Moreover, maximum pixels' coverage area formula can be deduced as in

$$A_{\text{np}} = w' \times h', \tag{3}$$

where all nonzero pixels are covered by the area where $w' \leq w$ and $h' \leq h$. In this equation, A_{np} is the number of nonzero pixels area in the tongue image after clustering, w is width area of tongue image before clustering, and h is the height area of tongue image before clustering. Samples of the deep red clusters and red/light red clusters identified by our proposed clustering identifiers are as shown in Figure 3.

During the accumulation of average colour of every clusters in k-means procedure, the average colour of the maximum pixels' coverage area clusters is observed to have the least concentration of red (denoted by low a^* and b^* attributes) in *Lab* colour compared to other clusters. Hence, after series of training procedures in SVM, this identifier has the best accuracy to classify deep red tongue.

3.2.3. Reduction of Support Vectors in SVM via Tongue Clustering Identifiers. This section describes the theory and fundamentals of our proposed SVM model aided with clustering identifiers to reduce the number of support vectors during training and testing procedures that will lead to a fast classification system with high accuracy. A support vector machine (SVM) is a supervised machine learning method that is defined by a separating hyper plane which can be used for classification of images. Given a set of labelled training data, the algorithm outputs an optimal hyper plane which predicts the new example to fall on which side of the gap. In the image processing concept, this training algorithm of SVM builds a model of mapping pixels and assigns them into one category or the other divided by a discriminative hyper plane. A good separation is achieved by the hyper plane that has the largest margin which describes the distance to the nearest training data point (support vectors) of any class. The larger the margin, the lower the generalization error of the classifier will be. The reason why SVM insists on finding the maximum margin hyper planes is that this optimization offers the best generalization ability. In other words, it compromises better classification performance (e.g., accuracy) on the training data of the future data. In addition, SVM can also perform a nonlinear classification using the kernel method by mapping their inputs into high-dimensional feature spaces implicitly. Besides, SVM is said to have high generalization ability for classification problem compared to other machine learning algorithms even though the input space is very high [36].



(a)

(b)

(c)

Figure 3: (a) Deep red cluster identified by maximum pixels' coverage area identifier and (b) and (c) red/light red clusters identified by maximum colour distance identifier.

Currently, there are two types of approaches for multiclass SVM. The first method is by considering all the data using one optimization formula and the other one is by combining several binary classifiers together and finally opt for the one with the best optimization. According to this paper [37], the formulation to solve multiclass SVM problems in one step requires the number of variables to be proportional to the number of classes. Therefore, there must be several binary classifiers to be constructed or a larger optimization problem is needed. Hence, in general, it is computationally more expensive to solve a multiclass problem than a binary problem with similar number of training databases [38]. Nevertheless, when the training data becomes bigger, more time is needed for optimization. In [39], one of the limitations of SVM classifier is its computational inefficiency when there are millions of instances to be classified. However, this problem can be solved by breaking the core sets into a series of subsets. Through this motivation, we have proposed a two-stage classifier that can break a large optimization problem into a series of small problems (clusters) where each problem only involves a small number of informative pixels (small number of support vectors too) so that the optimization problems can be solved successfully. Solving these small variables will definitely save its response time whilst the generalization ability can be improved significantly.

```
width=200; height=200;
DataSet = cell([], 1);
%list and returns a character vector containing the full path to the file
for i=1:length(dir(fullfile(Dataset2,'*.png')))
     % Training set process
     k = dir(fullfile(Dataset2,'*.png'))
%Reading image and its information
     for j=1:length(k)
        tempImage = imread(horzcat(Dataset2,filesep,k{j}))
        imgInfo = imfinfo(horzcat(Dataset2,filesep,k{j}))
        %Image conversion to grayscale and using intensity as classification parameter in SVM
        if strcmp(imgInfo.ColorType,'grayscale')
           DataSet{j} = double(imresize(tempImage,[width height]))
        else
      DataSet{j}= double(imresize(rgb2gray(tempImage),[width height])); %we only use the colour intensity
of grayscale picture to classify the images, we did measure the classification using hue (chromatic pixels),
but tongue intensity provide better classifications.
```

Algorithm 1: Pseudocodes of training input setting in SVM.

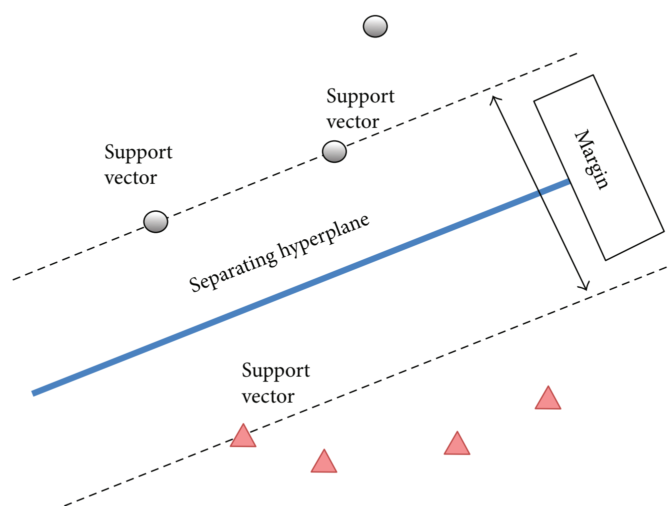


Figure 4: SVM concept of classification by constructed hyper plane.

Theoretically, the general binary (two classes) classification using SVM can be visualized in Figure 5. The motivation of our proposed algorithm is to have a margin that is as wide as possible to separate the training points between the two classes (deep red and light/red tongue) depending on the location of support vectors. The general equation of these hyper planes that separate these two categories can be described as

$$y_i\left(w^T x_i - b\right) \geq 1 \quad 1 \leq i \leq n$$
$$\min \frac{1}{2}\|w\|^2 \text{ s.t. } y_i\left(w^T x_i - b\right) \geq 1, \quad i = 1, 2, \ldots, n, \tag{4}$$

where w is the normal vector of the hyperplane, b is the bias, and x_i is the support vectors that lie on a feature space. In our case, the training input to SVM can be devised in pseudocodes given by Algorithm 1.

Methodically, if we apply this (4) to the set of test data, there will be many hyper planes generated to classify the data with respect to two different constraints. The best choice of hyper planes is the one that has the largest separation between the two classes as illustrated in Figure 4. These w and b parameters are solved to determine the classifier. The maximum margin can be determined by those x_i which lie nearest to it. These x_i are called boundary points or support vectors. The complexity of the SVM algorithm is really depending on the number of support vectors. In other words, the cost function does not explicitly depend on the dimensionality of feature space and the number of training samples. In soft margin cases, we need more training samples to get better generalization ability and lesser number of support vectors. In SVM, there are many types of kernel such as linear, Gaussian, quadratic, and polynomial. The kernel is defined as the inner product in the feature space. Kernel trick in SVM learns linear decision boundary in a high dimension

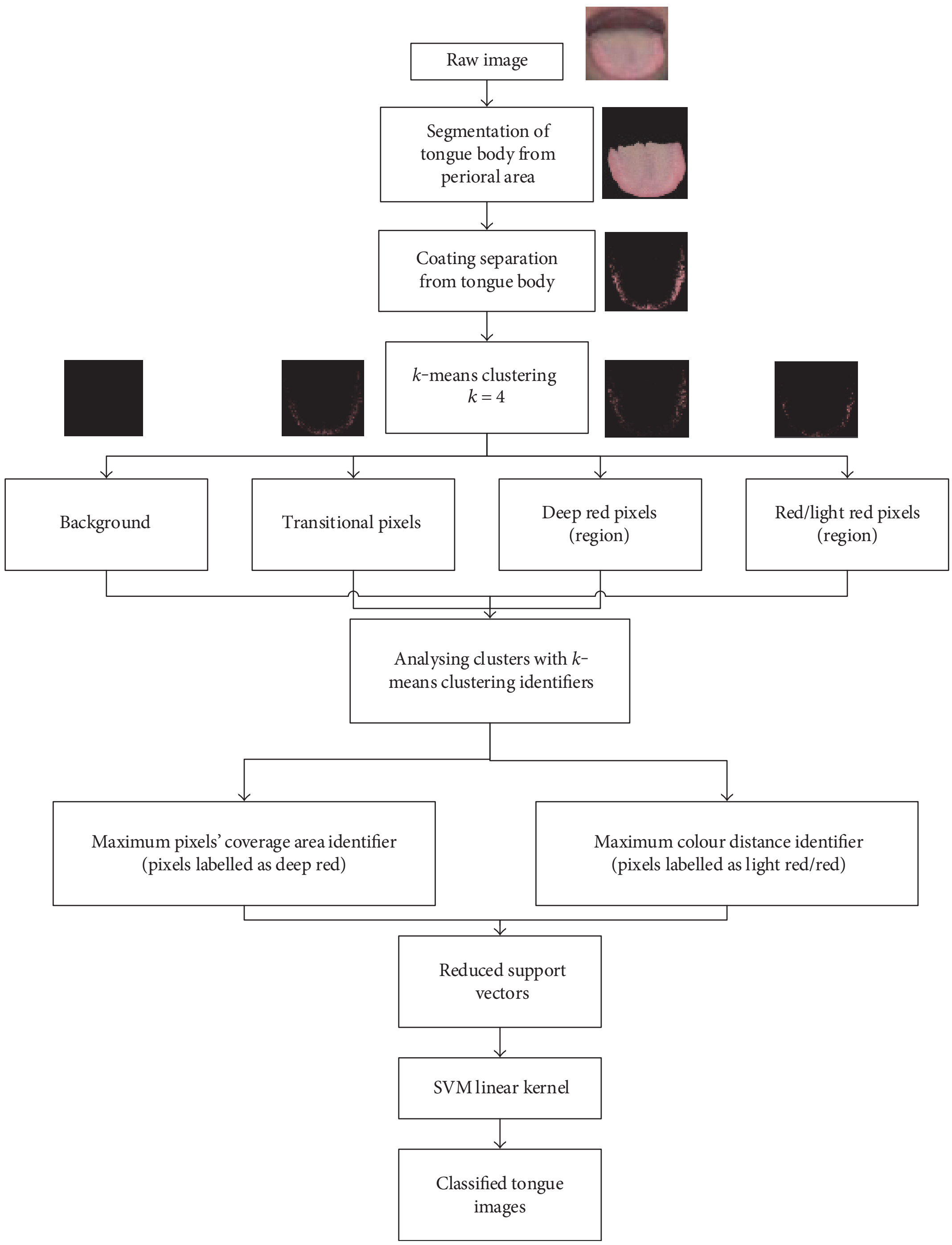


Figure 5: Flowchart of our proposed first-stage classification.

space without explicitly working on the mapped data. The linear kernel is the simplest example of a kernel and is obtained by considering the identity mapping for the feature space to satisfy $\varphi(x) = x$ or in formulation $k(x, x') = (x^T, x')$.

The down sampling procedures or sample selection method for training purposes before classification is essential to reduce classifier's burden and complexity. Since the tongue colour is very narrow [40], there are many overlapping pixels of similar colour exist. In other words, the distributions of tongue image pixels are usually nonlinear and continuous. In our proposed work, we have discovered that the most contributing colour and area analysis on the tongue by

Table 1: Red colour range for red and light red tongues.

Tongue colour	*Lab* colour range L^*	a^*	b^*
Red	$L^* < 56$	$32 \le a^* \le 39$	$6 \le b^* \le 10$
Light red	$L^* \ge 56$	$23 \le a^* \le 27$	$15 \le b^* \le 19$

```
I = lightred/red test_image_from_SVM;
I1 = I(:,:,1); % L
II1 = mean(nonzeros(I1));
I2 = I(:,:,2); % a
II2 = mean(nonzeros(I2));
I3 = I(:,:,3); % b
II3 = mean(nonzeros(I3));
%For b colour range checking:
colour_average = II3
if (colour_average>=6)&&(colour_average<=10)
  disp('Tongue is red colour');
else if (colour_average>=15)&&(colour_average<=19)
    disp('Tongue is light red colour');
   else
       disp ('Check b attributes');
       %repeat to a and L attributes
```

Algorithm 2: Pseudocodes of red colour ranges.

k-means clustering can be used to extract meaningful and informative colour region on the tongue body so that the redundant pixels can be eliminated. However, not all clusters produced by our proposed k-means clustering algorithm are relevant to be fed as training examples in SVM. To this motivation, we have developed two tongue clustering identifiers to be fed as training examples in SVM and they are called as maximum pixels' coverage area identifier and maximum colour distance identifier. An outline of our first-stage classification system aided by these clustering identifiers as training examples to distinguish the deep red and red/light red group of tongue can be visualized in Figure 5.

After series of training procedures, we have discovered that maximum colour distance identifier has the best ability to classify deep red tongue and maximum pixels' coverage area identifier has the best ability to classify red/light red tongue based on the loss function formula calculated using labelled tongue colour image databases. By using these proposed clustering identifiers, the numbers of overlapping pixels and misclassified points or outliers between the boundaries have been greatly reduced; hence, the number of support vectors is also reduced. This reduction of overlapping pixels promotes distance maximization (margin maximization) of separating hyper plane for better generalization ability. The implementation of our proposed tongue identifier as training examples is significant as it also lead to minimization of outliers that can prevent over fitting during the classification process. Moreover, SVM treats all training points equally; hence, both the noisy points and outliers will have negative impacts on the accurate classification [41]. Methodically, in order to classify some training points using an SVM model, the dot product of each support vector has to be computed with every test point. In other words, the SVM model did faster classification with fewer number of support vectors and vice versa. Thus, the computational complexity of our proposed classifier model aided with the clustering identifier model has been reduced via the reduction in the number of support vectors. The detailed measurement analyses and results of our proposed classification method will be discussed in the Experimental Results and Discussion. For SVM accuracy measurement, we have estimated the classification rate terminology via the loss function formula given by

$$\mathrm{ACC} \cong \left(\frac{n_{\text{red/light red}} + n_{\text{deep red}}}{N}\right) \times 100, \tag{5}$$

where ACC is the estimated successful classification accuracy rate, $n_{\text{red/light red}}$ is the number of red/light red tongue images that have been correctly classified, $n_{\text{deep red}}$ is the number of deep red tongue images that have been correctly classified, and N is the total number of tongue images used in classification.

3.3. Red Colour Range in Lab Colour Space. This section describes the second stage of classification where red colour range is used as a final classifier between red/light red groups of tongue after first classification is done. If the new example of tongue is classified as deep red tongue after the first stage of classification, then it will not be classified further using this red colour range. Nevertheless, if the new example is classified as red/light red tongue in the first stage of classification, then, it has to be classified further using red colour range for final verification. The measurement of red colour range is done during the clustering procedures where we have accumulated hundreds of average red and light red tongue colour clusters which are labelled clinically beforehand by the practitioners' naked eye as red and light red tongues. By naked eye, red and light red tongues look very similar because the colour range of chromatic value (a^* and b^* attributes) is relatively similar for red and light red tongues [33], but technically, the value of luminance attributes (L^*) is distinguishable. The red colour range in *Lab* colour space is summarized in Table 1.

After the first-stage classification, red/light red cluster will be tested using red colour range defined in Table 1. The new example of red/light red cluster should satisfy the range value of every attribute (L^*, a^*, b^*) in the table above. Red/light red

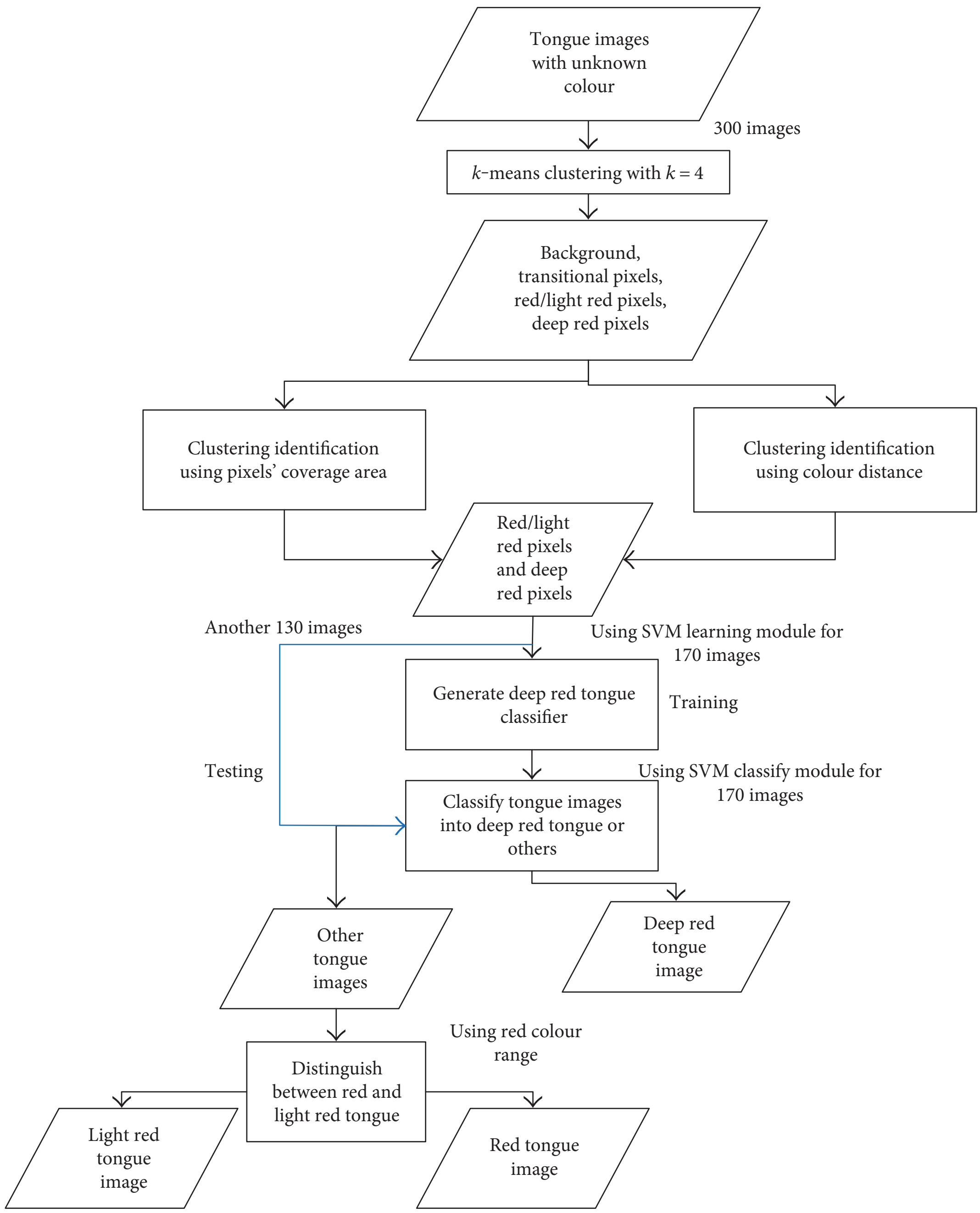


Figure 6: The outline of proposed computerized tongue colour diagnosis system.

cluster which does not fit the range of a^*, b^* value in Table 1 will be further classified based on luminance, L^* value afterwards. Subsequently, the user will get the final result of the tongue colour whether it is red or light red tongue. The red colour range pseudocodes were given by Algorithm 2.

The detailed procedures of our proposed tongue colour diagnosis system aided with the proposed identifiers and red colour range are illustrated in Figure 6.

4. Experimental Results and Discussion

4.1. Performance Analysis. This section discusses the experimental results and the performances of our proposed two-stage classification system using SVM model with clustering identifiers and red colour range. We have involved over 300 images and 600 features of our clustering identifiers as training examples in SVM to diagnose the red, light red, and deep red tongue colours automatically. These tongue images were taken by tongue image analyzing system (TIAS) on hundreds of outpatients in the Oriental Medicine Research Centre, Kitasato University in Japan, and each of the tongue colour was validated and labelled beforehand by several medical practitioners. The SVM algorithm is run in MATLAB environment with Intel® Core™ i7-3820CPU @3.60GHz. As a comparison, we have implemented raw images after segmentation and coating removal as training

Table 2: Comparison of average classification accuracy and execution time of several algorithms using same database specifications.

Method	Technique/kernel	Accuracy (%)	Execution time (s)
Conventional SVM (only SVM) [11, 13, 27, 28, 42]	RBF	50	219
	Polynomial	50	187
	Linear	57	249
	Quadratic	74	187
Proposed SVM with clustering identifiers (SVM + k-means)	RBF	63	166
	Polynomial	50	172
	Linear	89	149
	Quadratic	50	151
Neural network [18, 24]	Conventional	70.6	652

examples without clustering identifiers. Several types of kernels such as linear, Gaussian radial basis function (RBF), quadratic, and polynomial kernels have been tested. However, without clustering identifier, the accuracy rate measured was not convincing. The rows labelled as "Conventional SVM" in Table 2 summarizes the performance of the SVM algorithm tested with different kernels in terms of their classification accuracy and execution time. The last row in Table 2 shows the performance of neural network algorithm in classifying tongue colours whereby only 70.6% best classification rate was achieved.

Methodically, we have divided the experiment into two stages, SVM classifier and classification based on *Lab* colour ranges. The SVM is used to discriminate between the deep red and red/light red tongue colours aided by the proposed clustering identifiers. The best estimated classification rate attained is 89% when the linear kernel is being used as summarized in the rows labelled as "SVM with clustering identifiers" in Table 2.

Based on the measurement result by SVM on 300 images, we have deduced that linear kernel is the best kernel that can successfully separate the light red/red and deep red tongues using combination of maximum colour distance and pixels' coverage area identifiers. By implementing the result of k-means clustering algorithm that are based on colour distance and pixel's coverage area, the information of training and testing images to be tested by SVM had been simplified. Redundant pixels inside a tongue image after segmentation and coating removal that are not useful have been filtered out via our proposed clustering identifiers. By using these identifiers, there is a significant difference in classification boundary between these two classes (red/light red and deep red). In other words, the solution of SVM has been simplified.

However, a low classification accuracy have been observed in light red and red tongue classification by using maximum colour distance identifier in SVM because these two colours look very similar to the naked eye. Nevertheless, during the colour analysis measurement, the difference in light red and red tongues was observed via each of the *Lab* colour attributes: L^*, a^*, and b^* statistically. Therefore, in order to devise a reliable and precise diagnosis system, we have derived the red colour range from *Lab* colour space for red and light red tongues to be used to distinguish these two colours in second-stage classification such that overall classification accuracy is high for the whole tongue colour diagnosis system. These colour attributes were accumulated during the clustering procedures. Table 3 showed the comparison of accuracy between chromatic and luminance attributes between red and light red tongues. This chromatic and luminance attributes pattern indicator (*Lab* attributes) does not apply to deep red tongue because deep red tongue can be recognized by the first-stage classifier.

As can be seen in Table 3, low accuracy was recorded if only chromatic or colour attribute range (range for a^* and b^* only) was used to recognize the light red and red tongue colours. In the light red and red tongues, several pixel values of a^* and b^* are very similar; hence, there are some overlapping pixels between these two colours. Therefore, luminance can distinguish the light red and red more accurately. Here, chromatic with luminance attributes from the *Lab* colour space gave the highest accuracy in red and light red tongue second-stage classification which is 95%. In total, the estimated classification accuracy using our proposed two-stage classifier was recorded as 94%. Therefore, we have developed the intelligent diagnosis system that can predict and diagnose the tongue based on its colour and also can be an assisted tool for the practitioner during the clinical practices.

5. Conclusion

In this work, we have proposed a two-stage classification method to diagnose three tongue colours: red, light red, and deep red. The proposed automatic colour diagnosis system is very useful for early detection of imbalances condition inside the body. According to traditional medicine perspectives, light red tongue is considered normal; red tongue is always associated to excess of heat, dehydration, hemoconcentration, or irritability; and deep red tongue is associated with blood stagnation, coldness, and so forth. The first-stage classifier is mainly based on SVM aided by k-means clustering identifiers: maximum colour distance identifier and maximum pixel's coverage area identifier. These two identifiers have been employed to discriminate between the red/light red with deep red tongue colour with measured classification accuracy of around 89%. To further obtain the separation between the light red tongue and red tongue accurately, red colour range using *Lab* colour space were

Table 3: Comparison of red colour range's performance in classification.

Lab colour space attributes	Accuracy
Only chromatic attribute range (a^*, b^*)	63%
Both chromatic and luminance attribute range (L^*, a^*, b^*)	95%

introduced for red and light red clusters. These colour attributes were measured using average cluster value of red and light red clusters during clustering process. The accuracy of second-stage classification using red colour range has been recorded as 95%. Finally, using our proposed two-stage classifier, the overall estimated successful classification rate to discriminate red, light red, and deep red tongue colours is 94%. The whole algorithm execution time is around 48 seconds to diagnose one tongue image which offers fast processing time for online diagnosis. This intelligent tongue colour diagnosis system can be employed as an assisted tool for the practitioners and medical doctors to diagnose any unknown tongue colour image during their practice.

Conflicts of Interest

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

Acknowledgments

This study was supported by the Japan Society for the Promotion of Science KAKENHI Grant no. 26860420 from the Ministry of Education, Culture, Sports, Science and Technology, Japan; Kitasato University Research Grant for Young Researchers; and FRGS research grant vote project (4F504) from the Ministry of Higher Education, Malaysia.

References

[1] Y. Q. Liu, Y. X. Wang, N. N. Shi, X. J. Han, and A. P. Lu, "Current situation of International Organization for Standardization/Technical Committee 249 international standards of traditional Chinese medicine," *Chinese Journal of Integrative Medicine*, vol. 2016, pp. 1–5, 2016.

[2] T. Kawanabe, N. D. Kamarudin, C. Y. Ooi et al., "Quantification of tongue colour using machine learning in Kampo medicine," *European Journal of Integrative Medicine*, vol. 8, no. 6, pp. 932–941, 2016.

[3] X. Wang, B. Zhang, Z. Yang, H. Wang, and D. Zhang, "Statistical analysis of tongue images for feature extraction and diagnostics," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5336–5347, 2013.

[4] X. Wang and D. Zhang, "An optimized tongue image color correction scheme," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 6, pp. 1355–1364, 2010.

[5] I. G. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, vol. 160, Amsterdam, Netherlands, 2007.

[6] Z. Hao, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: discriminative nearest neighbor classification for visual category recognition," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2126–2136, Washington D. C., USA, June 2006.

[7] J. Wang, X. Wu, and C. Zhang, "Support vector machines based on K-means clustering for real-time business intelligence systems," *International Journal of Business Intelligence and Data Mining*, vol. 1, no. 1, pp. 54–64, 2005.

[8] Y. Yao, Y. Liu, Y. Yu et al., "K-SVM: an effective SVM algorithm based on K-means clustering," *JCP*, vol. 8, no. 10, pp. 2632–2639, 2013.

[9] Q. Gu and J. Han, "Clustered support vector machines," *AISTATS*, pp. 307–315, 2013.

[10] B. Pang, D. Zhang, N. Li, and K. Wang, "Computerized tongue diagnosis based on Bayesian networks," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 10, pp. 1803–1810, 2004.

[11] S. C. Hui, Y. He, and D. T. C. Thach, "Machine learning for tongue diagnosis," in *6th International Conference on Information, Communications & Signal Processing, 2007*, pp. 1–5, Singapore, December 2007.

[12] Y. Jiao, X. Zhang, L. Zhuo, M. Chen, and K. Wang, "Tongue image classification based on Universum SVM," in *2010 3rd International Conference on Biomedical Engineering and Informatics*, pp. 657–660, Yantai, China, October 2010.

[13] B. Zhang, X. Wang, J. You, and D. Zhang, "Tongue color analysis for medical application," *Evidence-Based Complementary and Alternative Medicine*, vol. 2013, Article ID 264742, p. 11, 2013.

[14] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.

[15] B. Pang, D. Zhang, and K. Wang, "Tongue image analysis for appendicitis diagnosis," *Information Sciences*, vol. 175, no. 3, pp. 160–176, 2005.

[16] Q. Li and Z. Liu, "Tongue color analysis and discrimination based on hyperspectral images," *Computerized Medical Imaging and Graphics*, vol. 33, no. 3, pp. 217–221, 2009.

[17] Q. Li, Y. Wang, H. Liu, Y. Guan, and L. Xu, "Sublingual vein extraction algorithm based on hyperspectral tongue imaging technology," *Computerized Medical Imaging and Graphics*, vol. 35, no. 3, pp. 179–185, 2011.

[18] L. Zhi, D. Zhang, J.-Q. Yan, Q.-L. Li, and Q.-L. Tang, "Classification of hyperspectral medical tongue images for tongue diagnosis," *Computerized Medical Imaging and Graphics*, vol. 31, no. 8, pp. 672–678, 2007.

[19] B. Zhang, B. V. Kumar, and D. Zhang, "Detecting diabetes mellitus and nonproliferative diabetic retinopathy using tongue color, texture, and geometry features," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 2, pp. 491–501, 2014.

[20] F. Haris, K. Purnama, and M. Purnomo, "Tongue image analysis for diabetes mellitus diagnosis based on SOM Kohonen," *ETBHL*, vol. 13, pp. 374–380, 2011.

[21] L. Lo and P. Hsu, "P02.93. Evaluation of blood stasis on tongue diagnosis associated with diabetes mellitus," *BMC Complementary and Alternative Medicine*, vol. 12, Supplement 1, p. 149, 2012.

[22] G. Liu, J. Xu, and D. Gao, "A method for tongue coat classification based on neural network integration [J]," *Computer Engineering*, vol. 14, p. 039, 2003.

[23] W. Aimin, Z. Zhongxu, and S. Lansun, “Research on the tongue color classification in automatic tongue analysis of traditional Chinese medicine,” *Beijing Biomedical Engineering*, vol. 3, p. 001, 2000.

[24] J. Jang, J. Kim, K. Park, S. Park, Y. Chang, and B. Kim, “Development of the digital tongue inspection system with image analysis,” in *Engineering in Medicine and Biology, 2002, 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002, Proceedings of the Second Joint*, pp. 1033–1034, Houston, TX, USA, October 2002.

[25] H. Zhang, K. Wang, D. Zhang, B. Pang, and B. Huang, “Computer aided tongue diagnosis system,” *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 7, pp. 6754–6757, 2005.

[26] Z. Xinfeng, X. Xiaozhao, and C. Yiheng, “Tongue image classification based on the TSVM,” in *2nd International Congress on Image and Signal Processing, 2009 (CISP '09)*, pp. 1–4, Tianjin, China, October 2009.

[27] R. Kanawong, T. Obafemi-Ajayi, T. Ma, D. Xu, S. Li, and Y. Duan, “Automated tongue feature extraction for ZHENG classification in traditional Chinese medicine,” *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 912852, p. 14, 2012.

[28] Z. Gao, M. Cui, and G. Lu, “A novel computerized system for tongue diagnosis,” in *International Seminar on Future Information Technology and Management Engineering, 2008 (FITME '08)*, pp. 364–367, Leicestershire, UK, November 2008.

[29] S. Haykin, *Neural Networks*, Prentice Hall, Upper Saddle River, NJ edition, 1999.

[30] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[31] S. Amari and S. Wu, “Improving support vector machine classifiers by modifying kernel functions,” *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[32] H. Byun and S.-W. Lee, “Applications of support vector machines for pattern recognition: a survey,” in *Pattern Recognition with Support Vector Machines*, pp. 213–236, Springer, Berlin, Heidelberg, 2002.

[33] N. D. Kamarudin, C. Y. Ooi, T. Kawanabe, and X. Mi, *Tongue's Substance and Coating Recognition Analysis Using HSV Color Threshold in Tongue Diagnosis*, p. 100110J-100110J-5, 2016.

[34] M. Sansone, R. Fusco, A. Pepino, and C. Sansone, “Electrocardiogram pattern recognition and analysis based on artificial neural networks and support vector machines: a review,” *Journal of Healthcare Engineering*, vol. 4, no. 4, pp. 465–504, 2013.

[35] P. Chen, *Diagnosis in Traditional Chinese Medicine*, Paradigm Publications, Brookline, MA, USA, 2004.

[36] O. Chapelle, P. Haffner, and V. N. Vapnik, “Support vector machines for histogram-based image classification,” *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055–1064, 1999.

[37] R. Debnath, N. Takahide, and H. Takahashi, “A decision based one-against-one method for multi-class support vector machine,” *Pattern Analysis and Applications*, vol. 7, no. 2, pp. 164–175, 2004.

[38] H. Chih-Wei and L. Chih-Jen, “A comparison of methods for multiclass support vector machines,” *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[39] X. Wu, V. Kumar, J. Ross Quinlan et al., “Top 10 algorithms in data mining,” *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[40] L. Zhuo, J. Zhang, P. Dong, Y. Zhao, and B. Peng, “An SA–GA–BP neural network-based color correction algorithm for TCM tongue images,” *Neurocomputing*, vol. 134, pp. 111–116, 2014.

[41] Y. Zhang, X. Zhang, W. Liu et al., “Automatic sleep staging using multi-dimensional feature extraction and multi-kernel fuzzy support vector machine,” *Journal of Healthcare Engineering*, vol. 5, no. 4, pp. 505–520, 2014.

[42] L. P. Zhong Gao, W. Jiang, X. Zhao, and H. Dong, “A novel computerized method based on support vector machine for tongue diagnosis,” in *2007 International Conference on Signal-Image Technology & Internet-Based Systems*, pp. 849–854, Shanghai, China, 2007.