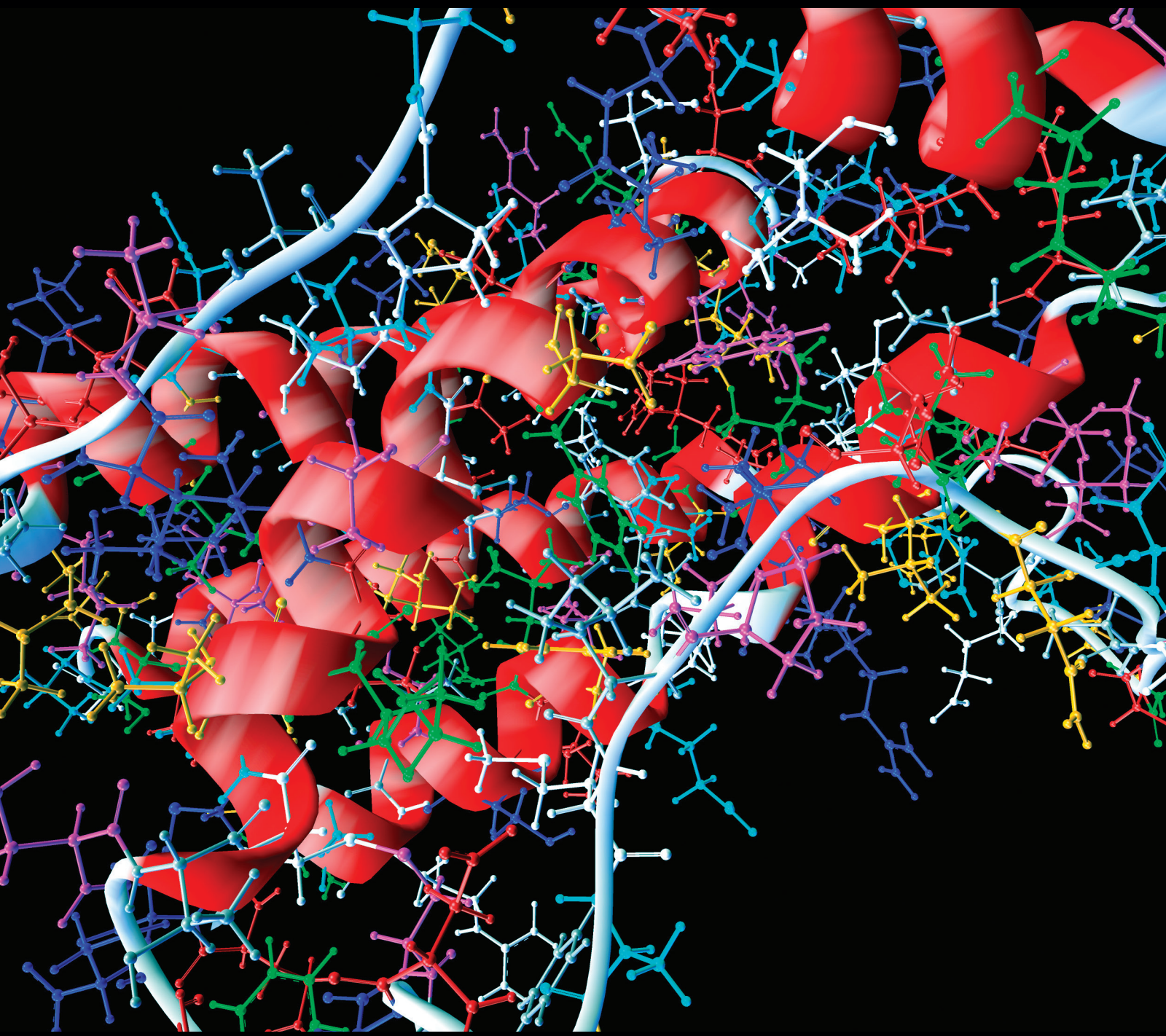
Computational and Mathematical Methods in Medicine

Machine Learning Applications in Medical Image Analysis

Guest Editors: Ayman El-Baz, Georgy Gimel'farb, and Kenji Suzuki



Machine Learning Applications in Medical Image Analysis

Computational and Mathematical Methods in Medicine

Machine Learning Applications in Medical Image Analysis

Guest Editors: Ayman El-Baz, Georgy Gimel'farb, and Kenji Suzuki

Editorial Board

Contents

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 2361061, 2 pages
https://doi.org/10.1155/2017/2361061

Editorial

Machine Learning Applications in Medical Image Analysis

Ayman El-Baz,[1] Georgy Gimel'farb,[2] and Kenji Suzuki[3]

[1]*Bioengineering Department, The University of Louisville, Louisville, KY, USA*
[2]*Department of Computer Science, The University of Auckland, Auckland, New Zealand*
[3]*Department of Radiology, The University of Chicago, Chicago, IL, USA*

Correspondence should be addressed to Ayman El-Baz; aselba01@louisville.edu

Received 4 April 2017; Accepted 4 April 2017; Published 13 April 2017

Multiple today's medical imaging modalities, for example, X-ray CT, MRI/fMRI, and PET scanners, supply computer-aided diagnostics (CAD) with a host of complex and highly informative images. The resulting big volumes of raw visual information are extremely difficult to handle. Thus new strategies for imaging-based CAD and therapies of diseases have to be developed.

In recent years, machine learning became one of the major tools of medical image analysis in various CAD applications. Prior knowledge being learnt from characteristic examples provided by medical experts helps to guide image registration, fusion, segmentation, and other computations towards accurate descriptions of the initial data and extraction of reliable diagnostic cues to reach the CAD goals. Inspired by and combined artificial intelligence, pattern recognition, biology, mathematical statistics, optimization, and many other fields of science, machine learning is successfully employed to find hidden relationships in the complex image data and link them to the goal diagnoses or monitoring of diseases. For a very simple example, learning quantitative 3D shape descriptors of the *corpus callosum* on brain MRI helps much in organizing a successful early CAD of autism or dyslexia.

This special issue pursues the goals of discussing challenges, technologies, and applications of machine learning in the present CAD. Careful reviewing of more than 31 submissions resulted in the selection of 12 papers covering the following topics: measuring topological DWI tractography to detect Alzheimer's disease; 3D kidneysegmentation from abdominal images; driver fatigue detection based on a single EEG channel; accuracy assessment for iterative closest point (ICP) registration; texture and morphological analyses of multiple regions of interest (ROI) to classify breast ultrasound (BUS) images; pulmonary nodule classification with deep convolutional neural networks; combined lung nodule classification with local difference patterns; automatic lung segmentation from thoracic CT; instrument detection and pose estimation in retinal microsurgery; deep and transfer learning for colonic polyp classification; research on techniques of multifeatures extraction for tongue image and its application in retrieval; and active learning to classify diabetic retinopathy.

N. Amoroso et al. used multiplex network concepts to characterize the brain organization from a topological perspective.

F. Khalifa et al. integrated discriminative features from current and prior visual appearance models into a random forest classifier to automatically segment 3D kidneys from dynamic CT images.

J. Hu combined four entropy features and ten classifiers to detect driver fatigue by processing an EEG.

G. Krell et al. compared different unconstrained ICP algorithms on realistic noisy data from an optical sensor of the tomotherapy HD system.

M. I. Daoud et al. combined multiple-ROI morphological and texture analyses to effectively segment BUS images.

W. Li et al. designed deep convolutional neural networks (CNNs) with strong autolearning and generalization abilities to classify lung nodules.

K. Mao and Z. Deng proposed local difference patterns (LDP) and combined classifiers to specify lung nodules on low-dose CT images.

J. Wang and H. Guo presented a fully automatic three-stage lung segmentation by skin boundary detection, rough determination of a lung contour, and pulmonary parenchyma refinement.

M. Alsheakhali et al. modeled detection, tracking, and pose estimation of a retinal microsurgical instrument as a conditional random field (CRF) inference in order to localize the instrument's forceps tips and center point and estimate the orientation of its shaft.

E. Ribeiro et al. explored automated classification of colonic polyps by deep learning of different pretrained or built from scratch trainable CNNs on 8-HD-endoscopic databases acquired by various imaging modalities.

L. Chen et al. presented a novel approach to extract color and texture features of tongue images. Results showed that the developed approach can improve the detection rate of lesion in tongue image relative to single feature retrieval.

Y. Zhang and M. An used an active learning based classifier of features extracted by recognizing anatomical parts and detecting lesions to identify retinal images and further reduce costs of screening the diabetic retinopathy.

Acknowledgments

We would like to thank the aforementioned contributors to this special issue, as well as reviewers for their hard and timely work. Finally, we give special thanks to the editorial board of this journal for their confidence in great machine learning potentialities in application to medical image analysis.

Ayman El-Baz
Georgy Gimel'farb
Kenji Suzuki

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 8064743, 11 pages
https://doi.org/10.1155/2017/8064743

Research Article

Research on Techniques of Multifeatures Extraction for Tongue Image and Its Application in Retrieval

Liyan Chen, Beizhan Wang, Zhihong Zhang, Fan Lin, and Yihan Ma

Software School of Xiamen University, Xiamen 361005, China

Correspondence should be addressed to Liyan Chen; chenliyan@xmu.edu.cn and Zhihong Zhang; zhihong@xmu.edu.cn

Received 3 August 2016; Revised 7 October 2016; Accepted 15 February 2017; Published 30 March 2017

Academic Editor: Ayman El-Baz

Tongue diagnosis is one of the important methods in the Chinese traditional medicine. Doctors can judge the disease's situation by observing patient's tongue color and texture. This paper presents a novel approach to extract color and texture features of tongue images. First, we use improved GLA (Generalized Lloyd Algorithm) to extract the main color of tongue image. Considering that the color feature cannot fully express tongue image information, the paper analyzes tongue edge's texture features and proposes an algorithm to extract them. Then, we integrate the two features in retrieval by different weight. Experimental results show that the proposed method can improve the detection rate of lesion in tongue image relative to single feature retrieval.

1. Introduction

Chinese traditional medicine is the experience of the Chinese people during thousand years of struggle with the disease. Its curative effect is significant, and side effects are small. Compared with modern medicine, it has certain advantages and potential in healthcare, health, rehabilitation, and so forth. In the Chinese traditional medicine, there are four kinds of diagnosis methods including inspection, olfaction, interrogation, and palpation. Tongue image is an important part of inspection which gets the disease's situation by observing patient's tongue color and edge shape changes [1].

Tongue diagnosis is one of the important topics in the field of medicine at present; with the continuous deepening of the Chinese traditional medicine tongue diagnosis objectiveness research, digital images of tongue diagnosis have also been applied in the clinical work. A lot of tongue images are generated in clinical works every day, and how to retrieve and manage the increasingly large tongue database to support tongue diagnosis features extraction has become a very challenging subject. Traditional tongue images management describes the tongue image information by manual labeling and retrieves tongue image by the description information, but this way has been unable to meet the needs of large-scale tongue image database retrieval. On the other hand, traditional tongue diagnosis depends highly on clinicians' experience and thus different clinicians are likely to reach remarkably different diagnostic results for the same patient. So this paper proposes a novel method applying content-based image retrieval technology to tongue image retrieval.

In recent years, some computer image processing technologies have been used in tongue diagnosis in the Chinese traditional medicine. These methods can be divided into two categories according to different tongue image features used: color-based approaches and texture-based ones. For the former, [2–7] used the color features to analyze the tongue images. The color matching of tongue images in different color spaces with different metrics was investigated and reported in [2, 4] that proposed a method based on region partition and feature matching for color recognition of tongue images. Li and Yuen (2002) addressed the problem of color image matching in medical diagnosis. They proposed the sorted metric in coordinate space. To improve the matching performance, a probabilistic combined metric is proposed based on the theory of combining classifier. Wang et al. (2004) proposed a new tongue color calibration scheme and utilized a gradient vector flow (GVF) snake based model integrating the chromatic information of the tongue image used to

extract the tongue body. Li and Liu (2009) developed a push broom hyperspectral tongue imager and discussed its spectral response calibration method. A new approach to analyze tongue color based on spectra with spectral angle mapper is presented. This new color analysis approach is superior to the traditional method especially in achieving meaningful areas of substances and coatings of tongue. Papers [8, 9] used a variety of tongue image features (such as color, texture, or shape) to match and identify tongue images. Chiu (2000) built a computerized tongue examination system (CTES) based on chromatic and textural algorithm. The chromatic algorithm is developed to identify the colors of the tongue and the thickness of its coating. Guo (2008) proposed a new color texture operator, Primary Difference Signal Local Binary Pattern. The matching performance is evaluated on color, grayscale and color texture, and fusion of color and texture features.

For these methods, each of them has its fair share of success, but corresponding limitations also accompany with them. Taken as a whole, they fail to satisfy the demands for both accuracy and robustness simultaneously, which are the basic requirements for a successful extraction. In fact, the Chinese traditional medicine attaches great importance to correlation analysis; namely, different tongues reflect different diseases, and there may exist some symbiosis or mutual exclusivity among different characteristics of the tongue. It is necessary for us to integrate a variety of tongue image features for tongue image analysis. Based on this, the paper proposes a method combining color and texture features in retrieval by different weight to improve the recognition rate for tongue diagnosis in the Chinese traditional medicine. First, we use the iterative method to extract the initial main colors and the number of them and then get the main color histogram by GLA algorithm. Considering the deficiency of expressing tongue image information by using color feature only, the paper analyzes tongue image further and puts an algorithm to get tongue image's texture feature. The new method combines the improved main color histogram descriptor and edge histogram descriptor to give different weights to the comprehensive retrieval.

To evaluate the performance of the proposed algorithm, in the course of the experiment, we used 268 tongue images as experimental samples; these images are divided into several colors in advance by the doctor.

From experimental results, we can see that the improved main color histogram algorithm is better than the traditional main color histogram in the search results. Although differences of the statistical results are small, we can obviously feel the effect of the retrieval greatly improved. At the same time, the position of the relevant tongue is also more forward and focused.

Experiment used the same tongue images ditto, which are divided into 5 texture categories. Experimental results show that improved edge histogram's retrieval precision and recall ratio is slightly higher than that of the traditional edge histogram.

Finally, in order to analyze the function of comprehensive color and texture feature in tongue image retrieval, we select a set of tongue images which have prominent color and texture features for the experiment and then randomly select a tongue image as the query image. In order to achieve better retrieval results, the weight of color and texture features is set to 0.6 and 0.4, respectively. We can see that the result is more accurate than the single feature search results. A large number of tongue image retrieval experiments show that, due to the great difference of the tongue images, using color or texture features in retrieval would be better for some tongue images.

Experiments show that new method can improve the detection rate of lesion in tongue images.

The paper is organized as follows. Section 2 reviews some extraction methods of tongue image features. An improved main color histogram method and improved edge histogram for tongue diagnosis in the Chinese traditional medicine are proposed in Section 3. The experimental results and analysis are shown in Section 4. Finally, some conclusions are drawn in Section 5.

2. Related Work

2.1. Main Color Descriptor. Color is the basic element of tongue image and it is one of the main features for tongue image recognition. Each tongue image has its own unique color feature, which is the basic and important feature in the image. Tongue and coating color links to the body contact, representing different lesions. So the color is one of the mainstays of tongue diagnosis in the Chinese traditional medicine and has important diagnostic value.

According to the theory of visual psychology [9], human's perception to image focuses on a few representative colors, while ignoring the secondary color details. MPEG-7 provides the main color descriptor to describe the main color information of image in arbitrary irregular region, which reflects the main color of the image. The main color descriptor is used as the color feature for image retrieval, and the basic idea is as follows [10].

To image I, first convert the color space to N dimension; then, the image color can be represented by an N dimension vector:

$$F = \{\{C_i, P_i\}\}, \tag{1}$$

where $i \in [0, N-1]$, C_i represents the image color after quantifying, and P_i represents a percentage of the corresponding quantitative color for the whole image. Sorting F by descending order, the traditional method is to use $P_i \geq 5\%$ as the main color.

GLA algorithm is an iterative clustering algorithm searching the optimal vector quantizer for target; it is an iterative split and union process. Algorithm 1 describes the GLA algorithm step [11].

2.2. Edge Histogram Descriptor. In the process of the Chinese traditional medicine, discriminating the tongue color and analyzing the lingual teeth marks, prick, crack, addiction, and other features are necessary; these features belong to the category of texture analysis. The edge texture's distribution is an important texture information and the edge histogram descriptor recommended by the MPEG-7 is widely used in

Input: Sample points for cluster n, Initial cluster center c_i
Output: Clusters k
Method:
(1) Input sample points for cluster n and Initial cluster center c_i.
(2) Repeat.
(3) Sort every sample points by the principle of proximity.
(4) Recalculate clusters center.
(5) Calculate the distortion value of all sample, until the distortion value is lower than the setting threshold.

Algorithm 1: GLA clustering algorithm (GLA algorithm).

the texture features of image retrieval based on the notion that, especially when the image texture distribution is not uniform, the descriptors' effect would be better when used for image matching [10].

Edge histogram describes five types of edge space distribution, containing four kinds of directional edges and a nondirectional edge. The basic idea of the algorithm is to divide the image into several subblocks and calculate the value of each subblock edge, depending on the direction of the subblock edge cumulative statistics and the edge histogram of the whole image, as follows:

(1) Divide the image into 4×4 subimages.

(2) Divide each subimage into smaller subimage blocks.

(3) According to the edge detection operator defined by MPEG-7, calculate five kinds of edge values of each subimage block. If the maximum edge value is greater than a certain threshold, set the direction as the edge direction of the image block.

(4) Get the 5 bin edge histogram of the subimages from the edge direction of image block and finally calculate the 80 bin histogram of the whole image.

Assuming that there are two images Q and T and their edge histograms are H_q and H_t, then Q and T can use the Minkowski formula where $r = 1$ to measure the similarity as follows:

$$D(Q,T) = \sum_{m=0}^{79} \left| h_q[m] - h_t[m] \right|. \tag{2}$$

3. Methods

The paper combines an improved MPEG-7 main color descriptor and improved edge histogram. They can overcome the problem of inaccurate retrieval by using single image feature, and they can improve the efficiency of retrieval. The two algorithms are introduced in detail in the following.

3.1. Improved Main Color Extract Algorithm. Each tongue image has its own unique color feature, which is the basic and important feature of the image. According to previous literature [8], there is no obvious difference between different kinds of RGB colors when using computer to automatically classify and identify tongue images in RGB color space. Therefore, RGB color space is difficult to represent the color features of different tongue images; this color space cannot classify the tongue image color. When using HSV color space to classify tongue image color, hue H angle value in turn increases according to the order of purple tongue, purple red tongue, pink tongue, yellow-coating tongue, pale tongue, and white-coating tongue's color feature; the saturation S value decreases and brightness V value increases in turn of purple red tongue, purple tongue, pink tongue, pale tongue, yellow-coating tongue, and white-coating tongue. This result can classify the tongue color. Therefore, it is necessary to use color features to convert the RGB color space into HSV color space while making the tongue image retrieval.

The HSV space can be obtained by the nonlinear changes of the RGB space; the conversion formula is as follows:

$$\begin{aligned} V &= \max(R,G,B), \\ S &= \frac{V - \min(R,G,B)}{V}. \end{aligned} \tag{3}$$

Set

$$\begin{aligned} r' &= \frac{V - R}{V - \min(R,G,B)}, \\ g' &= \frac{V - G}{V - \min(R,G,B)}, \\ b' &= \frac{V - B}{V - \min(R,G,B)}. \end{aligned} \tag{4}$$

Then,

$$H' = \begin{cases} 5 + b', & R = \max(R,G,B),\ G = \min(R,G,B) \\ 1 - g', & R = \max(R,G,B),\ G \neq \min(R,G,B) \\ 1 + r', & G = \max(R,G,B),\ B = \min(R,G,B) \\ 3 - b', & G = \max(R,G,B),\ B \neq \min(R,G,B) \\ 3 + g', & B = \max(R,G,B),\ R = \min(R,G,B) \\ 5 - r', & \text{other.} \end{cases} \tag{5}$$

$$H = 60 \times H'$$

We can know from the above formula that

$$H \in [0, 360^{\circ}], \quad S \in [0, 1], \quad V \in [0, 1]. \tag{6}$$

The main color of the image can be extracted by clustering method, and the selection of the initial cluster centers has great influence on the results of the image color classification in GLA algorithm. Because the division of the tongue image color is not obvious, the effect of the initial clustering center of the random selection is not good. In this section, we use an iterative method to determine the initial clustering number and clustering center and then use GLA algorithm to extract the main color histogram.

The specific algorithm process to determine the number of primary colors and the initial color is as follows:

(1) To specify color image $I(x, y)$, set its scale as $I(x, y)$; if color space is RGB space, convert it into HSV space according to formulae (3)–(5).

(2) Quantify the HSV space to nine subsections; the formula is as follows:

$$\text{area} = \begin{cases} 0, & v \le 0.2 \text{ (black)} \\ 1, & s \le 0.1,\ 0.2 < v \le 0.8 \text{ (gray)} \\ 2, & s \le 0.1,\ 0.9 < v \le 1 \text{ (white)} \\ 3 + h' \text{ (other)} \end{cases}$$

$$h' = \begin{cases} 0, & h \in (315, 360] \cup [0, 20] \text{ (red)} \\ 1, & h \in (20, 75] \text{ (yellow)} \\ 2, & h \in (75, 155] \text{ (green)} \\ 3, & h \in (155, 190] \text{ (cyan)} \\ 4, & h \in (190, 260] \text{ (blue)} \\ 5, & h \in (260, 315] \text{ (purple)}. \end{cases} \tag{7}$$

Scan image I and calculate the number of pixels belonging to the nine subspace $s_0, s_1, \ldots, s_8$ and the probability of the image $p_0, p_1, \ldots, p_8$, respectively.

(3) Set a threshold, calculate the interval of which $p_i > T$ number, and record the value of the space area, stored in the array MC[k]; in the experiment, choose T = 15%.

(4) The number of intervals k determined by step (3) is the number of the main colors. And MC[k]'s value area is just the approximate range of the main color but cannot be used as the main color of the image. For example, MC[k] = 4's color represents yellow, but it can be divided into dark yellow, light yellow, and so on. Therefore, we still need to continue to iteratively calculate the main color of image.

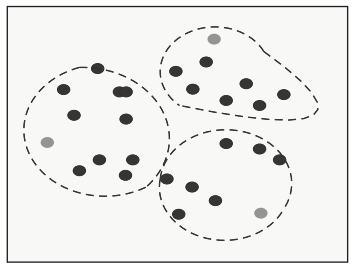


Figure 1: Clustering example graph of step (I).

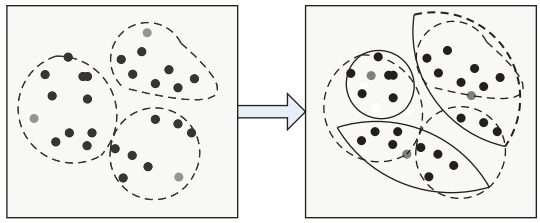


Figure 2: Clustering example graph of step (II).

The original main color MC[k] and the main color number k are used to obtain the main color by GLA algorithm, and then calculate the main color histogram; the steps are as follows:

(I) Classify each pixel $I'(j)$ in the image. According to formula (8), divide $I'(j)$ to interval where its pixel values are close to initial main color. ω_i is weighting coefficient. Figure 1 describes clustering example graph of step (I).

$$d_i = \sum \omega_i \left\| I'(j) - \text{MC}_i \right\|^2. \tag{8}$$

(II) Clarify cluster center. Recalculate every color interval's cluster center as new color after classifying pixels. n_i is the number of pixels in the MC_i interval. Figure 2 describes clustering example graph of step (II).

$$\text{MC}_i = \frac{\sum n_i I'(j)}{\sum n_i}. \tag{9}$$

(III) Repeat executed step (I) and step (II), until cluster center MC[k] does not change.

(IV) Perform the splitting operation. According to formula (10), calculate the errors between every color cluster interval, if error is greater than threshold T_1, dividing the color interval into two new color intervals and calculating the center of new interval, $\text{MC}_{\text{new1}} = \text{MC}_i - d/2, \text{MC}_{\text{new1}} = \text{MC}_i + d/2$. Repeat steps (I), (II), and (III). Figure 3 describes clustering example graph of step (IV).

$$d_i = \frac{1}{n_i} \sum \sqrt{\left(I'(j) - \text{MC}_i\right)^2}. \tag{10}$$

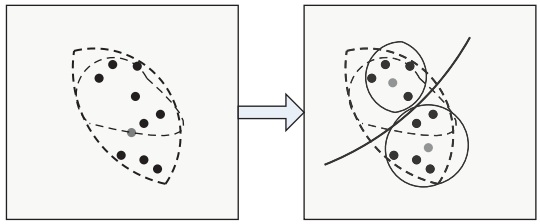


Figure 3: Clustering example graph of step (IV).

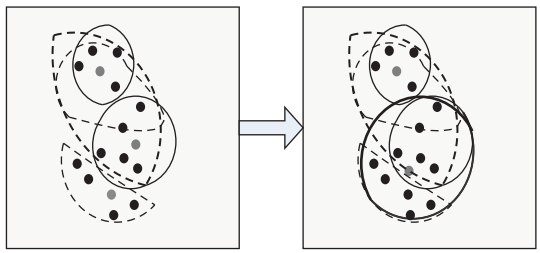


Figure 4: Clustering example graph of step (V).

(V) Perform merging operation. Calculate the distance of cluster centers, if the distance of two color cluster intervals is less than threshold T_2, unify the two interval. According to formula (11), the new clustering center is calculated. Repeat steps (I), (II), and (III). Figure 4 describes clustering example graph of step (V).

$$\mathrm{MC}_{\mathrm{new}} = \frac{\left(n_i \mathrm{MC}_i + n_j \mathrm{MC}_j\right)}{n_i} + n_j. \tag{11}$$

(VI) Perform the clustering end. While MC[k] does not change, divide or unify the cluster end.

To the $M * N$ color image $I(x, y)$, use algorithm in Section 3.1 to get image main color descriptor as follows:

$$F = \{(c_1, \mu_1), (c_2, \mu_2), \ldots, (c_k, \mu_k)\} = \left\{\left(c_1, \frac{n_1}{M * N}\right), \left(c_2, \frac{n_2}{M * N}\right), \ldots, \left(c_k, \frac{n_k}{M * N}\right)\right\}, \tag{12}$$

where c_i is the main color and μ_i presents the probability that c_i happen.

Figure 5 describes a tongue image and the main color histogram extracted from the algorithm.

3.2. Improved Edge Histogram Extract Algorithm. The ability of color feature to distinguish tongue images containing dry fur, crack, and other prick space prominent positions is not strong. Using texture features of tongue image edge to describe the tongue image retrieval can achieve better retrieval effect on spatial distribution of tongue image. In Section 2.2, the edge histogram extraction algorithm only describes the local edge information of the image and the improved algorithm adds a global edge histogram and a semiglobal edge histogram to make up the shortage.

The improved edge histogram extract method's steps are as follows:

(1) Set $I(x, y)$ to be a gray image whose scale is $M * N$; gray level is L; if I is color image in RGB space, use the following formula to convert color image into gray image:

$$g = 0.299 * r + 0.587 * g + 0.144 * b. \tag{13}$$

(2) Divide I into 4×4 subimages, on average, $I_1, \ldots, I_{16}$. Calculate every subimage's local edge histogram; every subimage contains 5 bin (0°, 90°, 45°, 135° and nondirection), so the whole image has 16×5 bin = 80 bin.

(3) Divide every subimage I_i into fixed number of image blocks; the area of image block changes as the area of the whole image. The number of image blocks in experiment is 256.

(4) Every subimage block can be seen as four 2×2 macroblocks; each of the macroblocks' edge detection operator in each direction is not the same. Calculate the five kinds of edge value of each image block and take the maximum value; if the maximum value is greater than the threshold value, then set the direction as the edge of the image block. Experimental results show that the best threshold is 20. Direction θ's edge values are calculated as follows:

$$E_\theta = \left|\sum_{i=0}^{3} a_i(m, n) * f_\theta(i)\right|, \tag{14}$$

where $a_i(m, n)$ represents the average gray value of the i macroblock and $f_\theta(i)$ represents i macroblocks' edge detection operator in direction θ. θ is 0°, 90°, 45°, 135°, and no direction.

(5) Get the subimage I_i's 5 bin edge histogram from 256 image blocks; the whole image of the edge histogram is 80 bin.

(6) Normalize and quantify the edge histogram got in step (5), and then with nonlinear quantified value 80 bin which has to be normalized, each histogram uses fixed 3 bit to encode reduced amount of computation.

(7) Global histogram represents the edge distribution information of whole image, calculated by adding and averaging the distribution information of the subimage in five directions; the dimension of the global histogram is 5. Set the local edge histogram of the image as EH; then the global histogram is

$$\mathrm{GH}_\theta = \sum_{i=1}^{16} \mathrm{EH}_\theta(i). \tag{15}$$

(8) Semiglobal histograms represent image region horizontal, vertical, and adjacent block edge information. As shown in Figures 3–11, 1~4 subblocks represent the

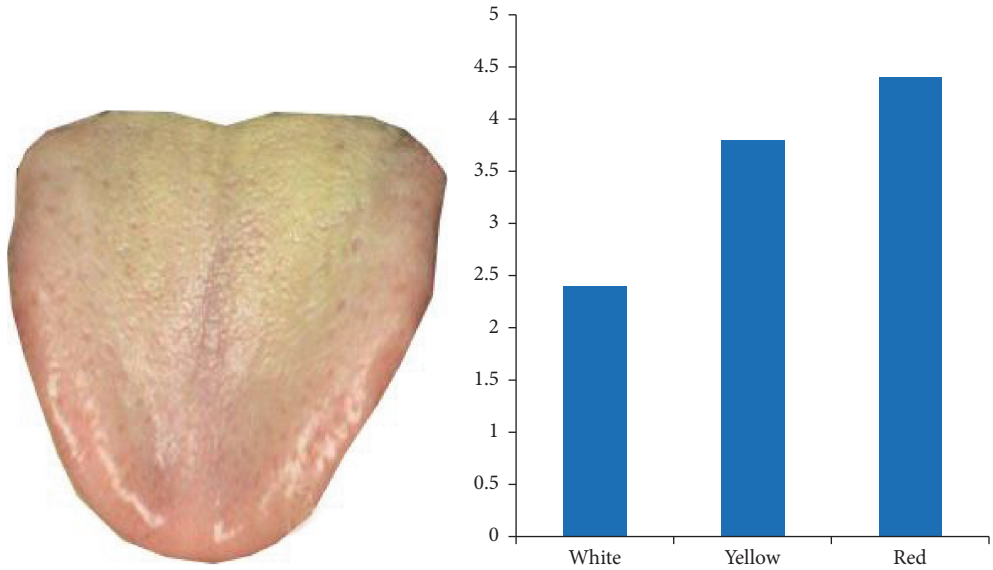


Figure 5: Tongue image and its main color histogram. (a) Original image. (b) Main color histogram.

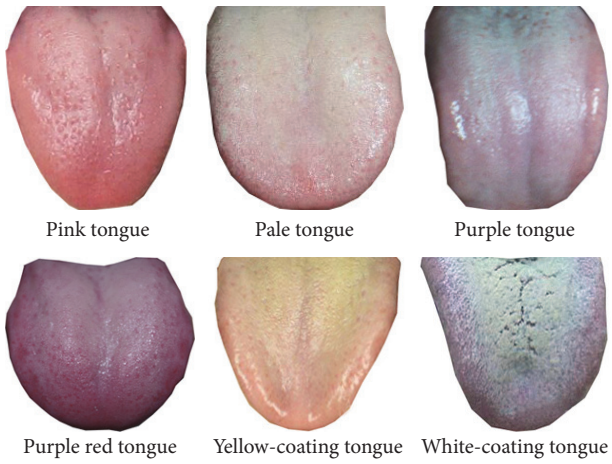


Figure 6: Six tongue image samples with Color Feature in database.

vertical edge information, 8 subblocks represent the horizontal edge information, 9–13 subblocks represent adjacent block edge information, and semiglobal histogram of the whole image is 13 × 5 bin = 65 bin.

3.3. Feature Extract. To the $M*N$ color image $I(x, y)$, use the algorithm in Section 3.1 to get image main color descriptor as follows:

$$F = \{(c_1, \mu_1), (c_2, \mu_2), \ldots, (c_k, \mu_k)\} = \left\{\left(c_1, \frac{n_1}{M*N}\right), \left(c_2, \frac{n_2}{M*N}\right), \ldots, \left(c_k, \frac{n_k}{M*N}\right)\right\}, \quad (16)$$

where c_i is the main color and μ_i represents the probability that c_i happen.

Get the image's local edge histogram by using algorithm in Section 3.2 EH $= \{E_{0^\circ}, E_{45^\circ}, E_{90^\circ}, E_{135^\circ}, E_{\text{non-direction}}\}$; global histogram GH $= \{g_1, g_2, \ldots, g_5\}$ and semiglobal histogram SGH $= \{S_{0^\circ}, S_{45^\circ}, S_{90^\circ}, S_{135^\circ}, S_{\text{non-direction}}\}$.

Consider $E_\theta = \{e_{\theta,1}, e_{\theta,2}, \ldots, e_{\theta,16}\}$, where $e_{\theta,j}$ represents the image edge value in j subimage of θ direction: $g_i = \sum_{j=1}^{16} e_{\theta_i,j}$, $\theta_1 = 0^\circ$, $\theta_2 = 45^\circ$, $\theta_3 = 90^\circ$, $\theta_4 = 135^\circ$, $\theta_5 =$ non-direction; $S_\theta = \{s_{\theta,1}, s_{\theta,2}, \ldots, s_{\theta,13}\}$, $s_{\theta,1}$ represents 13 subblocks standing for semiglobal information of image's edge value in θ direction.

3.4. Similarity Measurement. To a given image, the algorithm can extract the main color F, the local edge histogram EH, the global edge histogram GH, and the semiglobal edge histogram SGH, where EH is an 80-dimensional vector and SGH's dimension is 65.

Set the histograms EH_q, GH_q, SGH_q and EH_t, GH_t, SGH_t as Q and T's local edge histogram, global edge histogram, and semiglobal edge histogram. Adding the weight of the global histogram to increase the impact of the image, Q and T's texture similarity is defined as

$$D(Q, T) = \sum_{i=0}^{79} \left|\text{EH}_q[i] - \text{EH}_t[i]\right| + 5 \times \sum_{i=0}^{4} \left|\text{GH}_q[i] - \text{GH}_t[i]\right| + \sum_{i=0}^{64} \left|\text{SGH}_q[i] - \text{SGH}_t[i]\right|. \quad (17)$$

The proposed algorithm is a comprehensive retrieval for the color and texture of the image. If the distance between query image Q's main color histogram and target image T's

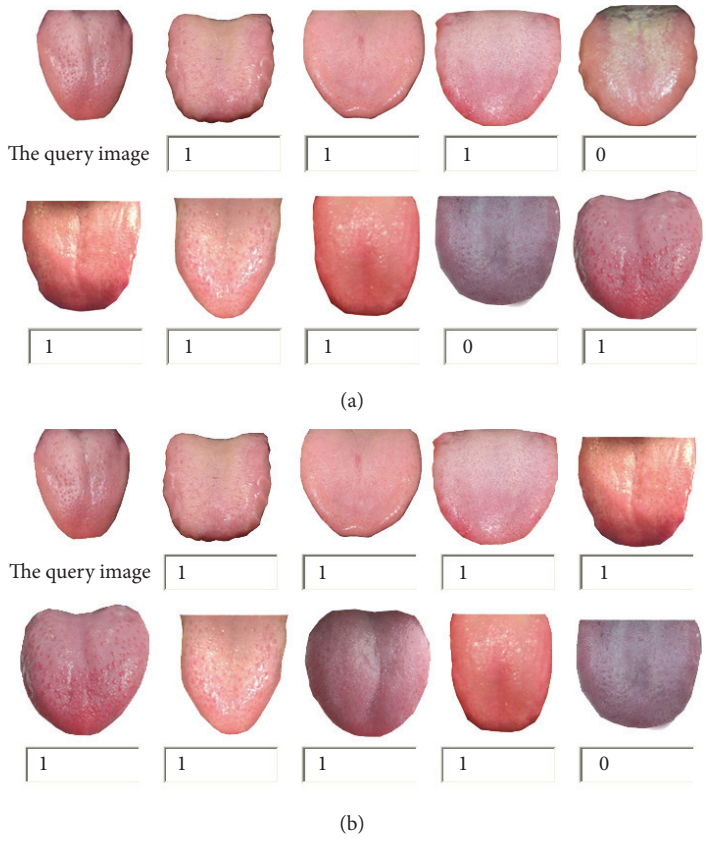


FIGURE 7: Results of two search algorithms based on color. (a) The top 9 results of tradition main color retrieval. (b) The top 9 results of improved main color histogram retrieval.

main color histogram is d_1, edge histogram gets the distance d_2, d_1's range is $[0, \max_1]$, and d_1's range is $[0, \max_2]$. The greater the distance value is, the more the two images are not similar. To make d_1 and d_2 able to be compared, normalize them: $d_1 = (\max_1 - d_1)/\max_1$; $d_2 = (\max_2 - d_2)/\max_2$.

d_1 and d_2's range after normalization is $[0, 1]$. If two images are the most similar, the similarity measurement is 1; otherwise, the least similarity measurement is 0 and the similarity measurement is a value of 0~1. While $d = \omega_1 d_1 + \omega_2 d_2$ retrieves the main color and edge histogram, using the distance of similarity, ω_1 represents the weight of main color and ω_2 represents the weight of texture feature. Generally speaking, the weight is 0.6 : 0.4. To different image information and practical applications, we can increase a certain weight to achieve better retrieval results.

4. Experiment Result and Analysis

4.1. Color Feature Retrieval Experiment. To evaluate the performance of Section 3.1 of the proposed algorithm, in Experiment 1, we used 268 tongue images as experimental samples; these images are divided into several colors in advance by the doctor. The total tongue images were divided into 6 categories, respectively, pink tongue, pale tongue, purple tongue, purple red tongue, yellow-coating tongue, and white-coating tongue, and each category contains at least 30 images. Six tongue image samples are shown in Figure 6.

We randomly selected a sample from each category as an example of tongue image and then retrieved it in the database. System first calculated the color feature vector and then similarity matched the color feature vector of the tongue image in the feature library. The similarity of the Euclidean distance is used in the paper, finally the returned tongue image is most similar to the sample.

Take pink tongue as an example, Figure 7 represents the first nine images based on the traditional main color algorithm and the improved algorithm of the main color.

Two kinds of algorithm's retrieval performance can be displayed from the retrieval system. In Figure 3, each set of images' upper left corner image is the image to be retrieved; the others are retrieval results. "1" is the related image and "0" is not related image. From experimental results, we can see that the improved main color histogram algorithm can usually be compared with the most similar images of those related images in advance, which is more consistent with the human visual perception.

To further compare the performance of the two algorithms, we, respectively, used traditional principal color algorithm and improved main color histogram algorithm to make a lot of tongue image retrieval experiments, calculated average precision of two algorithms in different tongue images, and then calculated the average precision of each algorithm to compare two algorithms' integrated retrieval performances.

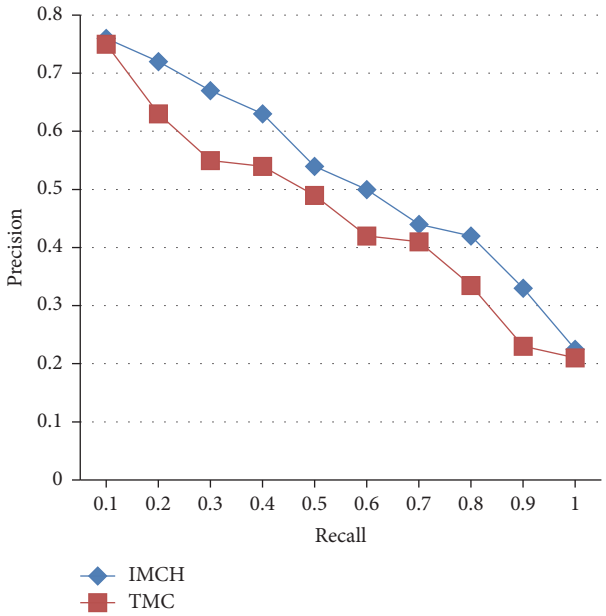


Figure 8: Average precision-recall curve of IMCH and TMC.

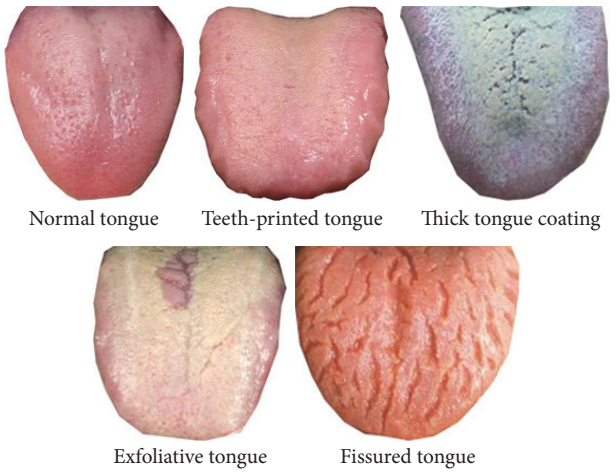


Figure 9: Six tongue image samples with Texture Feature in database.

Precision is defined as the ratio between the target images and the all images searched: precision = M/L, and recall is defined as the ratio between the target images in the result queue and the target images in the database: recall = M/D. Here L represents the total number of images returned by retrieval results, M represents the number of target images associated with the query image in the query results, and D represents the number of target images from image library, related to the image to be queried. The higher the precision is, the better the algorithm retrieval becomes.

Six groups of images are selected from tongue image database to build a retrieval set, forming 12 times retrieval. Figure 4 shows the precision comparing results between this paper's algorithm and tradition main color retrieval method. TMC represents the retrieval results based on tradition main color retrieval method, and IMCH represents the retrieval results based on the improved main color histogram method.

Experimental results show that the improved main color histogram algorithm is better than the traditional main color histogram in the search results. Although differences of the statistical results are small, we can obviously feel the effect of the retrieval greatly improved. At the same time, the position of the relevant tongue is also more forward and focused.

4.2. Texture Feature Retrieval Experiment. To evaluate the performance of Section 3.2 algorithms, experiment used the same tongue images ditto, which are divided into 5 categories: normal tongue, teeth-printed tongue, thick tongue coating, exfoliative tongue, and fissured tongue, and each category contains at least 30 images. Five tongue image samples are shown in Figure 9.

We randomly selected a sample from each category as an example of tongue image and then retrieved it in the database. The similarity of the Euclidean distance is used in this paper, and finally the returned tongue image is most similar to the sample.

Using fissured tongue image as an example, Figure 10 represents the retrieval results according to traditional edge histogram algorithm and improved edge histogram algorithm, the first nine images sorted according to the size of the similarity.

We repeatedly retrieved each type of tongue image. Figure 11 shows the precision and recall ratio comparing results between this paper's algorithm and tradition edge histogram retrieval method. TEH represents the retrieval results based on tradition edge histogram retrieval method, and IEH represents the retrieval results based on the improved edge histogram method.

Experimental results show that improved edge histogram's retrieval precision and recall ratio is slightly higher than that of the traditional edge histogram. Although differences of the statistical results are small, we can obviously feel the effect of the retrieval greatly improved. At the same time, the position of the relevant tongue is also more forward and focused.

4.3. Comprehensive Feature Retrieval Experiment. To analyze the function of comprehensive color and texture feature in tongue image retrieval, we select a set of tongue images which have prominent color and texture features for the experiment and then randomly select a tongue image as the query image. In order to achieve better retrieval results, the weight of color and texture features is set to 0.6 and 0.4, respectively.

Figure 12 represents the retrieval results of the main color algorithm, improved edge histogram algorithm, and image retrieval algorithm based on the main color and edge histogram, in accordance with the first nine images according to the size of the similarity.

Figure 12(a) is the result of only using color features. The similarity is gradually reduced from left to right, from top to bottom. Although the retrieval tongue image is similar to the query image in the color, the texture pattern of the last two images is obviously different from the query image.

(a) Retrieval based on tradition edge histogram

(b) Retrieval based on improved edge histogram

Figure 10: Results of two search algorithms based on texture. (a) The top 9 results of tradition edge histogram retrieval. (b) The top 9 results of improved edge histogram retrieval.

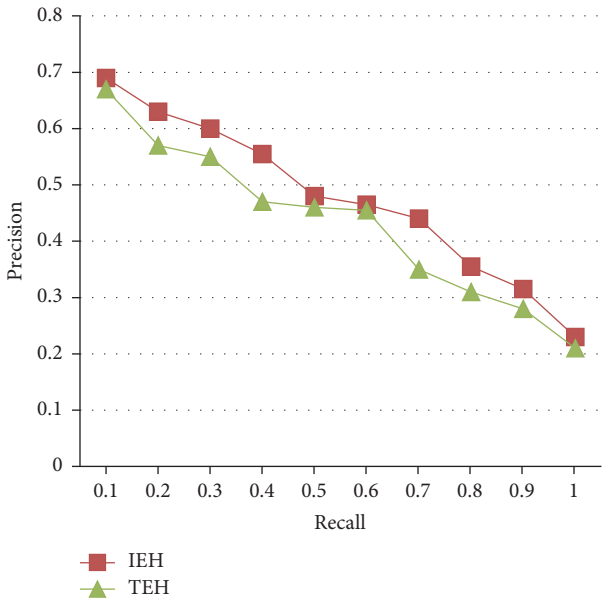


Figure 11: Average precision-recall curve of IEH and THE.

Figure 12(b) is the result of only using texture features. Although the retrieval performance is better than that in Figure 12(a), but it retrieved a tongue image with entirely different color. Figure 12(c) shows a comprehensive color and texture of the two features and for the same tongue image retrieval results; we can see that the result is more accurate than the single feature search results.

Table 1: Search result analysis.

Method	Precision	Recall
Improved main color histogram	0.4036	0.3481
Improved edge histogram	0.3983	0.2953
Comprehensive main color and edge histogram	0.5113	0.3620

To further compare the performances of the three algorithms, the paper uses a training set method, with 5 times cross validation, and the distribution feature of the training set is sufficient to describe the distribution feature of the entire image set. In this way, when adding new image to the training set, it will not affect the distribution feature of the entire image database and each image in training set is used in experiment as a query image. Calculate the average precision and recall ratio of the training set of images; the experimental results are shown in Table 1.

From Table 1, we can see that when using algorithm 1 to retrieve tongue images, precision effect is slightly higher than that of algorithm 2; the precision of algorithm 3 tongue image

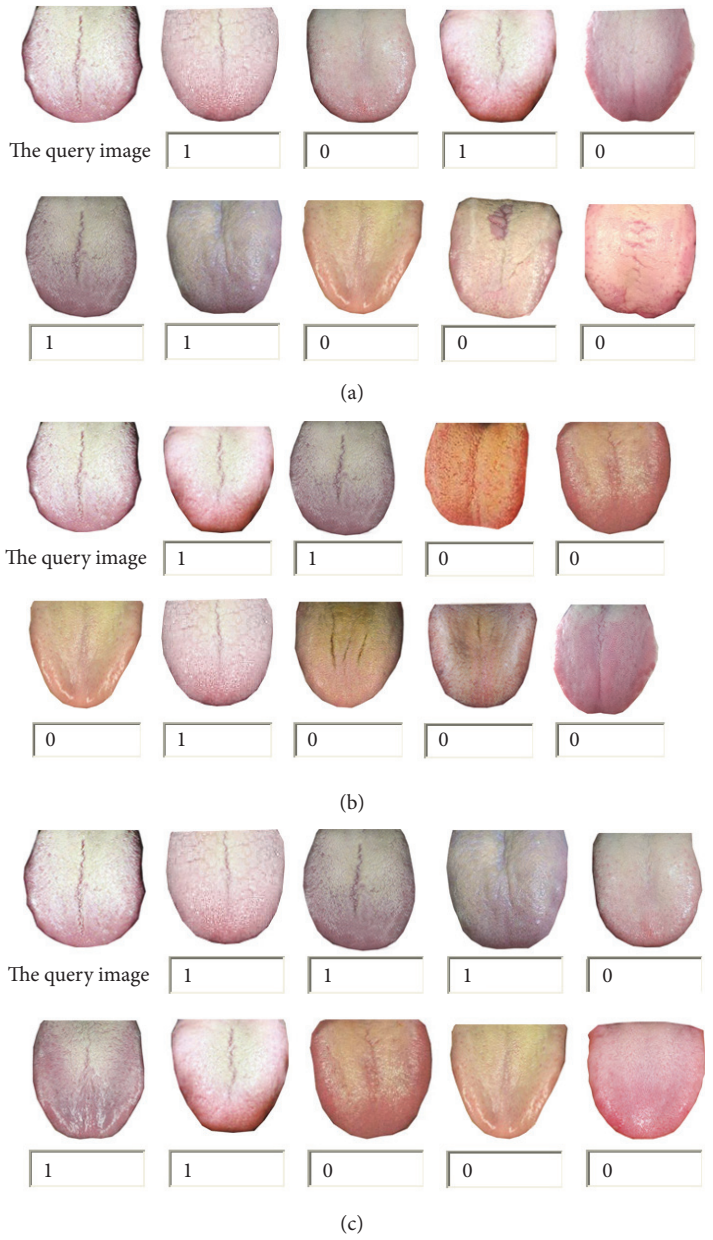


Figure 12: Results of three search algorithms based on color and texture. (a) The top 9 results of improved main color histogram retrieval. (b) The top 9 results of improved edge histogram retrieval. (c) The top 9 results of comprehensive main color and edge histogram retrieval.

retrieval is higher than that of the previous two algorithms, but the average recall ratio is low. The reason is the difference of the tongue images' color performance, which can be expressed better by the feature extraction of algorithm 1. Because the difference of tongue image texture feature is not such obvious, the extraction of the feature of algorithm 2 is difficult. Algorithm 3 can focus on both color and texture features, so it achieves a higher precision and recall ratio.

A large number of tongue image retrieval experiments show that, due to the great difference of the tongue images, using color or texture features in retrieval would be better for some tongue images. Therefore, to know tongue images in the practical application, we first judge the color and texture of the tongue image according to human's vision and then select the different retrieval methods and weights to obtain more satisfactory results.

5. Conclusion

The paper takes tongue image as an example; the research focuses on the key technology of image feature extraction and the technology research of the last layer which is the measurement of the similarity distance, so as to realize the content-based retrieval of the tongue image with a specific diagnostic value.

This paper first uses the iterative method to extract the initial main colors and the number of them and then gets

the main color histogram by GLA algorithm. Considering the deficiency of expressing tongue image information by using color feature, the paper analyzes tongue image further and puts an algorithm to get tongue image's texture feature. The new method combines the improved main color histogram descriptor and edge histogram descriptor to give different weights to the comprehensive retrieval. Experiments on the 268 images including normal tongue, teeth-printed tongue, thick tongue coating, exfoliative tongue, fissured tongue, and variety of colors verify the effectiveness and robustness of this method. Experiments show that new method can improve the detection rate of lesion in tongue images.

The content-based image retrieval technology has a certain practical significance in the Traditional Chinese Medicine. Using this technique, the information can be extracted directly from the tongue image database, which avoids the subjectivity of the manual annotation of the tongue image and greatly reduces the manual workload. The research according to this subject will have broad application prospects. The results of objectivity tongue diagnosis play a positive role in promoting the Chinese traditional medicine research. How to combine with clinicians stagnant standard to extract more high level feature and identify the lesion images aiming at the different manifestations of the disease is a subject that needs further research.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Fujian Province Education Sciences Foundation of China (no. JAT160010), the National Science and Technology Supporting Plan (no. 2015BAH55F05), the Fundamental Research Funds for the Central Universities in China (no. 20720160073), the Fujian Province Soft Sciences Foundation of China (no. 2014R0091), the National Natural Science Foundation of China (no. 61502402), and the Natural Science Foundation of Fujian Province of China (no. 2015J05129).

References

[1] T. T. Deng, "Basic theory of traditional Chinese medicine," in *Diagnostics of Chinese Medicine*, pp. 5–11, Chih-Yin Publishing, Taipei, Taiwan, 1995.

[2] C. H. Li and P. C. Yuen, "Tongue image matching using color content," *Pattern Recognition*, vol. 35, no. 2, pp. 407–419, 2002.

[3] Y. Wang, Y. Zhou, and J. Yang, "A tongue analysis system for tongue diagnosis in traditional Chinese medicine," in *Proceedings of the International Symposium Computational and Information Science (CIS '04)*, vol. 3314 of *Lecture Notes in Computer Science*, pp. 1181–1186, Springer, 2004.

[4] Y.-G. Wang, J. Yang, Y. Zhou, and Y.-Z. Wang, "Region partition and feature matching based color recognition of tongue image," *Pattern Recognition Letters*, vol. 28, no. 1, pp. 11–19, 2007.

[5] 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.

[6] M. F. Zhu and J. Q. Du, "A novel approach for color tongue image extraction based on random walk algorithm," *Applied Mechanics and Materials*, vol. 462-463, pp. 338–342, 2014.

[7] B. Pang, D. Zhang, and K. Wang, "Tongue image analysis for appendicitis diagnosis," *Information Sciences*, vol. 175, no. 3, pp. 160–176, 2005.

[8] C.-C. Chiu, "A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue," *Computer Methods and Programs in Biomedicine*, vol. 61, no. 2, pp. 77–89, 2000.

[9] Z. Guo, "Tongue image matching using color and texture," in *Proceedings of the International Conference on Medical Biometrics (ICMB '08)*, pp. 273–281, January 2008.

[10] M. Lantagne, M. Parizeau, and R. Bergev, "Vision tool for comparing images of people," in *Proceedings of the 16th IEEE Conference on Vision Interface*, June 2003.

[11] M. Ji, "MPEG-7 color, texture and shape descriptors," *Computer Engineering and Applications*, no. 26, pp. 46–47, 2004.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 5271627, 10 pages
https://doi.org/10.1155/2017/5271627

Research Article

Topological Measurements of DWI Tractography for Alzheimer's Disease Detection

Nicola Amoroso,[1,2] Alfonso Monaco,[2] Sabina Tangaro,[2] and Alzheimer's Disease Neuroimaging Initiative

[1] *Università degli Studi di Bari "A. Moro", Via Orabona 4, 70123 Bari, Italy*
[2] *Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Via Orabona 4, 70123 Bari, Italy*

Correspondence should be addressed to Nicola Amoroso; nicola.amoroso@ba.infn.it

Received 4 August 2016; Accepted 27 October 2016; Published 2 March 2017

Academic Editor: Ayman El-Baz

Neurodegenerative diseases affect brain morphology and connectivity, making complex networks a suitable tool to investigate and model their effects. Because of its stereotyped pattern Alzheimer's disease (AD) is a natural benchmark for the study of novel methodologies. Several studies have investigated the network centrality and segregation changes induced by AD, especially with a single subject approach. In this work, a holistic perspective based on the application of multiplex network concepts is introduced. We define and assess a diagnostic score to characterize the brain topology and measure the disease effects on a mixed cohort of 52 normal controls (NC) and 47 AD patients, from Alzheimer's Disease Neuroimaging Initiative (ADNI). The proposed topological score allows an accurate NC-AD classification: the average area under the curve (AUC) is 95% and the 95% confidence interval is 92%–99%. Besides, the combination of topological information and structural measures, such as the hippocampal volumes, was also investigated. Topology is able to capture the disease signature of AD and, as the methodology is general, it can find interesting applications to enhance our insight into disease with more heterogeneous patterns.

1. Introduction

Recent years have shown an increasing interest for graph-based measures in magnetic resonance imaging (MRI) and diffusion-weighted imaging (DWI) studies focused on brain diseases [1–6]. Among neurodegenerative diseases, Alzheimer's disease (AD) is the most common type of dementia affecting over 5 million people [7, 8] and is characterized by a well-known stereotyped pattern involving a whole brain left privileged atrophy, especially affecting some regions related to cognitive functionality as the hippocampus [9–13]. However, it is not clear yet whether the combined use of MRI and DWI modalities can significantly enhance its diagnosis.

Previous machine learning studies, investigating mixed cohorts of normal controls (NC) and AD patients, have reported conflicting results, an even more evident effect with the inclusion of mild cognitive impairment (MCI) subjects. In some cases the combination of DWI and MRI features reported a significant classification improvement [14, 15]; in others these results were not confirmed [16]. It is obvious that a fair comparison should require common data sets and validation techniques; nevertheless, it is manifest that a primary role is played by the different features adopted. Different features, in fact, not only provide a different base of knowledge (which naturally affects the machine learning models) but also capture different clinical aspects. Measures based on directional diffusion, such as fractional anisotropy (FA), have been extensively used as they are able to detect the connectivity impairment effect of AD [17]. Some studies revealed remarkable effects with axial and radial diffusivity (λ_1, RD) [18, 19]. In other cases huge effects were revealed in RD or mean diffusivity (MD) [20]. Finally, even if it is FA to be largely adopted, in some cases it can result in being insensitive [21, 22].

It is worth noting that the vast majority of reported results are focused on voxelwise DWI-related measures more than global connectivity metrics. However, the recent

developments of more accurate and sophisticated processing pipelines for tractography reconstruction [23, 24] have encouraged the exploration of connectivity and topological measures to quantify the brain changes [25, 26]. Typical findings especially inherent to AD are related to connectivity disruption, eventually characterized by a loss of small world [27, 28] or rich club organization of the brain [29, 30]. AD patients exhibit a decreased network efficiency, implying abnormal topological organization [31, 32].

These studies are based on two, not necessarily competing, underlying hypotheses; that is, brain dysfunctions can be yielded by (i) a local connectivity impairment [33] or by (ii) an abnormal overall organization of the brain [34, 35]. The local impairment hypothesis has been largely confirmed. However, for the second hypothesis encouraging results have been reported. Indeed, topological measures can have detectable effect size [36, 37].

A holistic approach which describes the AD effects from a topological perspective is adopted here. More than focusing on local impairments we look for discriminating patterns in the brain connectivity organization; thus, DWI tractography is used to introduce a diagnostic topological score. As for the chosen cohort T1 MRI scans were also available; the score is compared and combined with volumetric measures to assess its informative content. The presented methodology is general, even tested in this case on Alzheimer's disease. It allows a description of the overall brain topology; thus, its application to diseases with less stereotyped patterns [38], such as Schizophrenia or Multiple Sclerosis, could give further insight.

2. Materials and Methods

2.1. Brain Connectivity Matrices. Data used in the preparation of this article were obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

For the present study 99 subjects from Alzheimer's Disease Neuroimaging Initiative (ADNI) including normal controls (NC) and Alzheimer disease (AD) patients were analyzed. We chose this cohort in order to have for each subject both T1 MRI and DWI brain scans. T1-weighted sequences (voxel size = $1.2 \times 1.0 \times 1.0\,\text{mm}^3$; TI = 400 ms; TR = 6.98 ms; TE = 2.85 ms; flip angle = 11) and DWI scans (voxel size = $2.7 \times 2.7 \times 2.7\,\text{mm}^3$) are described in detail on the ADNI website (http://adni.loni.usc.edu/wp-content/uploads/2010/05/ADNI2_GE_3T_22.0_T2.pdf); in particular for DWI 46 separate images were acquired: 5 T2-weighted ($b = 0\,\text{s/mm}^2$ images) and 41 diffusion-weighted images ($b = 1000\,\text{s/mm}^2$). Demographics and clinical information are shown in Table 1.

Table 1: Data size, age range, gender, and a cognitive score (Mini Mental State Examination (MMSE)) are shown for each diagnostic group: normal control (NC) and Alzheimer's disease (AD) subjects. Mean and standard deviation are shown when appropriated.

	Size	Age	Gender	MMSE
NC	52	73 ± 6	M/F 26/28	29 ± 1
AD	47	75 ± 9	M/F 29/18	23 ± 2

For each subject DICOM images were acquired from ADNI database. MRICRON software was used to convert DICOM to NIFTI format, with the *dcm2nii* suite. Then FMRIB Software Library (FSL) by the Oxford Centre for Functional MRI of the Brain, and in particular its diffusion toolkit FDT, was used for the complete image processing pipeline; see Figure 1 for the overall flowchart:

(1) Eddy current correction was performed to mitigate artifacts caused by eddy currents in the gradient coils.

(2) Brain extraction was performed to erase nonbrain tissue from each subject scan, thus reducing the computational burden of the analysis.

(3) An affine registration of all scans was employed to spatially normalize the whole data set to the MNI152 template. With this step the image processing phase was concluded.

(4) Bayesian estimation of diffusion parameters and the inherent tensor fitting was obtained with sampling techniques at each voxel. This step was preparatory for running the subsequent probabilistic tractography.

(5) Finally, probabilistic tractography was performed to obtain the connectivity matrix of each subject. Specifically, the Harvard-Oxford cortical atlas (http://neuro.imm.dtu.dk/wiki/Harvard-Oxford_Atlas) was used, thus resulting in a brain parcellation of 96 regions, 48 per hemisphere.

The final output was a weighted symmetric connectivity matrix $\mathscr{W}$ whose elements w_{ij} represented the strength of connectivity, that is, the number of fibers, between the ith and jth regions. The fundamental step of the whole image processing was the fiber reconstruction. The FDT tool generates a probabilistic streamline or a sample from the distribution on the location of the true streamline. By taking many such samples the histogram of the posterior distribution on the streamline location or the connectivity distribution is then built. Finally, the most probable traits connecting two regions are computed. We averaged the traits connecting region i to j and vice versa j to i to obtain a symmetric matrix. We considered all non-null connections, disregarding the weight information and obtaining a binary connectivity matrix $\mathscr{C}$ whose elements c_{ij} were straightforwardly defined:

$$c_{ij} = \begin{cases} 1 & \text{if } w_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

As the focus of this study was the topological organization of the brain, we privileged the study of $\mathscr{C}$; nonetheless, we

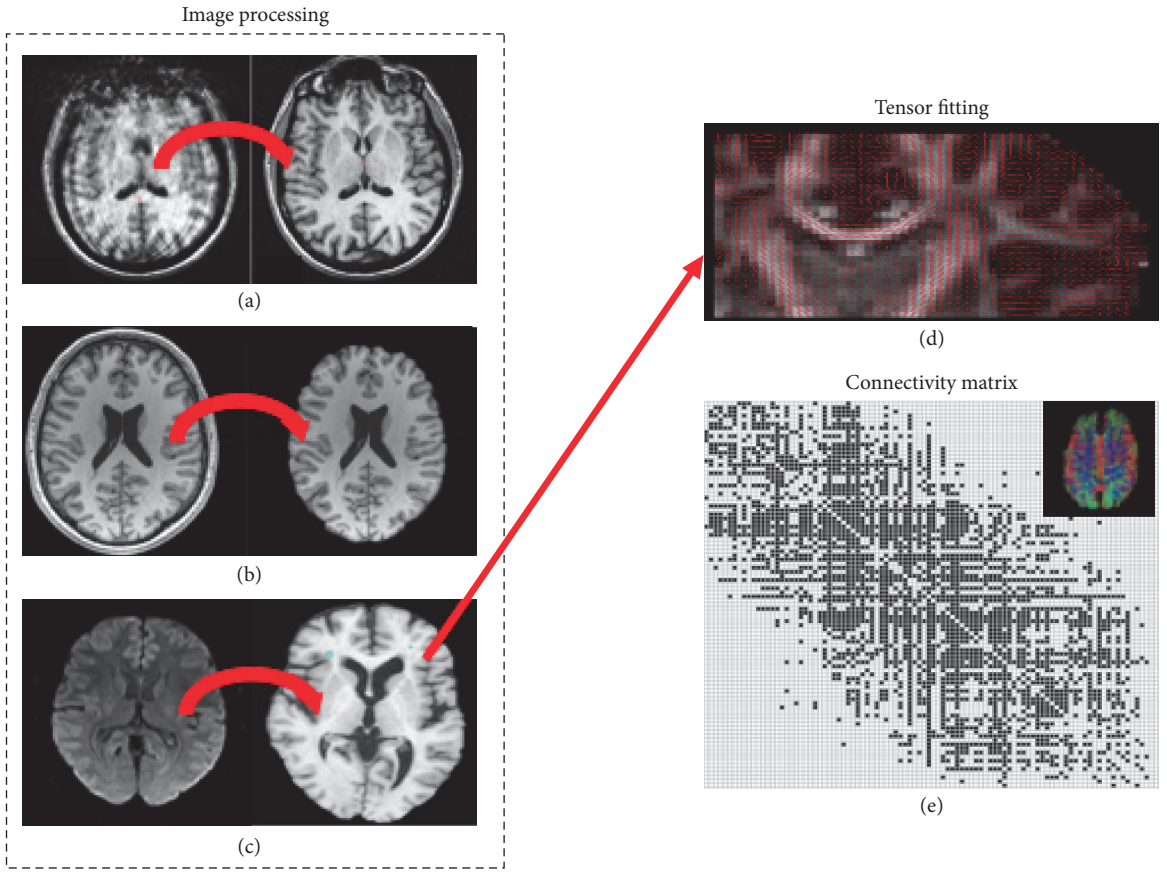


FIGURE 1: The figure shows the processing pipeline underwent by brain DWI scans. The dotted box includes the dedicated image processing steps: (a) eddy correction, (b) brain extraction, and (c) affine registration. For each voxel the diffusion tensor was estimated, (d) thus allowing the probabilistic fiber reconstruction. Using the Harvard-Oxford atlas, the connectivity matrix derived from tractography was computed for each subject.

also investigated the information carried by the connectivity weights. In principle, weight information should help the cohort discrimination as weights are directly affected by the impairment caused by the disease. However, it is worthwhile to note that tractography is very sensitive to artifacts and noise due to reconstruction algorithms and as a consequence it sometimes shows biological insights difficult to interpret.

2.2. Topological Overlap. The binary connectivity matrix $\mathscr{C}^{\alpha}$ of each subject α in the cohort is a compact representation of connected brain regions. A reasonable and partially confirmed hypothesis, deriving from the AD peculiarity of being a neurodegenerative disease, is that connectivity impairment should have a direct effect on the network topology. Besides, the impairment should reflect the severity of the pathological condition; thus, it should be expected that, for severe AD conditions, topology should manifest more evident changes. Nonetheless, natural biological variability can sometimes conceal these local effects and a huge statistical power will be required to investigate each brain connection and get a significant measurement.

We propose instead of describing the connectivity loss with a global indicator, trying to capture the whole brain behavior. To capture the whole informative content of the cohort in one comprehensive model we chose to adopt the novel multiplex network framework. A multiplex network, from now onward simply multiplex, is by definition a collection of networks sharing the same nodes [39]. Generally adopted in social sciences, this concept is naturally introduced to describe system with heterogeneous interactions. As an example, scientific authors with a common publication can be represented as a network; if this operation is stratified considering, for example, different journals or editors, a multiplex description arises. Another common example concerns the different relationships a group of people can share: social, geographical, and physical, just to mention a few.

The same concept applies here if we consider the anatomical districts as the fixed nodes of a network and build a network for each subject as if subjects were representing a stratification factor. Given a collection of these single subject networks, the multiplex can be visually represented as a 3D structure formed by M layers, one layer for each subject α, as shown in Figure 2.

Let $\boldsymbol{\mathscr{C}} = (\mathscr{C}^1, \ldots, \mathscr{C}^{\alpha}, \ldots, \mathscr{C}^{M})$ be the multiplex with each single subject graph $\mathscr{C}^{\alpha}$ formed by N nodes, the 96 labeled regions of the Harvard-Oxford atlas, and M layers (layers

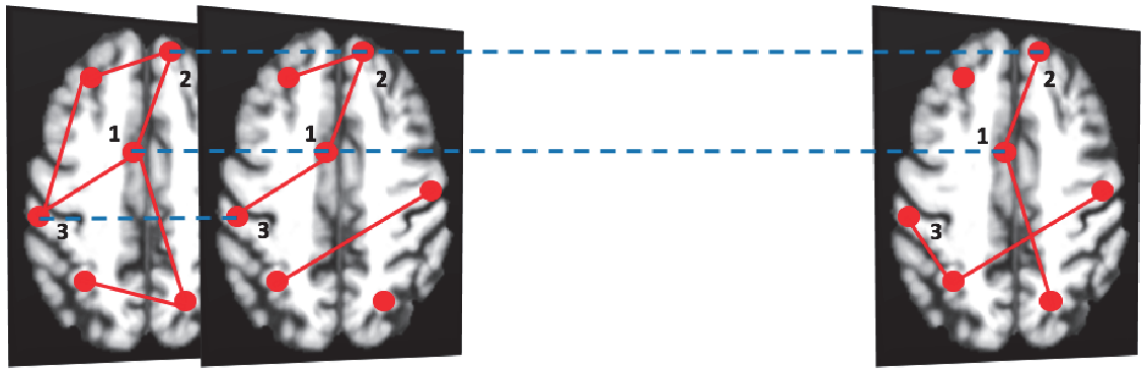


Figure 2: All subjects of the cohort are represented through graphs with exactly the same nodes (in red), corresponding to brain anatomical districts, but with different connections. For example, the figure shows the case of two nodes (1, 2) connected for all subjects and two nodes (1, 3) connected only in the first two subjects. The presence of a link in different subjects is also outlined (blue dashed lines). For each link in the networks it is possible to measure the fraction of subjects having a link in common, the so-called link overlap.

and subjects in this work will be interchangeably used). For a generic node of the multiplex i and for two generic subjects α and α' it is possible to define the local node overlap $n_i^{\alpha,\alpha'}$ [40] which is the total number of nodes j linked to the node i in a couple of layers α and α':

$$n_i^{\alpha\alpha'} = \sum_{j=1}^{N} c_{ij}^{\alpha} c_{ij}^{\alpha'}. \tag{2}$$

This is really useful information when investigating how central the node is within each layer, for example, to understand if there is a direct association between the kind of relationship defining the layer and the role played within it by a particular node.

However, from a topological point of view this is not very useful information, because what defines topology is not how intense the connections are, but their existence. Thus, adopting the same strategy to our case, we introduce here the link overlap matrix $\mathcal{O}$ and its elements o_{ij}:

$$o_{ij} = \frac{1}{M} \sum_{\alpha=1}^{M} c_{ij}^{\alpha}. \tag{3}$$

This matrix counts the number of times a link is present within each layer M. It is therefore a symmetric matrix whose values lie in the $[0, 1]$ interval.

It is reasonable to expect that link overlap should characterize important correlations among the different layers. One of the questions addressed by the present work is whether this measurement can detect the cross-sectional differences within a mixed NC/AD cohort. Accordingly, we built the multiplexes of NC subjects and AD patients. For both cases, the link overlap matrices $\mathcal{O}_{\mathcal{NC}}$ and $\mathcal{O}_{\mathcal{AD}}$ were computed. These matrices became binary with a 0.5 threshold for both NC and AD cohorts, considering it a likelihood measure assigned to each link.

The link overlap matrices represent the connectivity backbone of each population; in fact a qualitative difference can be directly observed by comparing the NC and the AD cases as shown Figure 3.

The overlap difference matrix $\mathcal{D}$ defined as

$$\mathcal{D} = \mathcal{O}_{\mathcal{NC}} - \mathcal{O}_{\mathcal{AD}} \tag{4}$$

has some interesting properties. It is a symmetric matrix whose elements $d_{i,j}$ are 0 for all connections with an identical behavior in both NC and AD cohorts and ± 1 for those connections present, respectively, only in $\mathcal{O}_{\mathcal{NC}}$ or $\mathcal{O}_{\mathcal{AD}}$. To emphasize these differences we introduce for each subject α a topological connectivity score $\mathcal{S}^{\alpha}$ as the Hadamard, that is, element-wise, product of $\mathcal{C}^{\alpha}$ and $\mathcal{D}$:

$$\mathcal{S}^{\alpha} = \sum_{i,j=1}^{N} \frac{1}{2} c_{ij}^{\alpha} d_{ij} \tag{5}$$

with d_{ij} representing the elements of $\mathcal{D}$ and the division by 2 takes into account the symmetry of $\mathcal{C}^{\alpha}$ and $\mathcal{D}$. In the same way a weighted connectivity score $\mathcal{S}_w^{\alpha}$ can be introduced by considering in the previous equation the original connectivity matrix $\mathcal{W}^{\alpha}$ and its elements w_{ij}:

$$\mathcal{S}_w^{\alpha} = \sum_{i,j=1}^{N} \frac{1}{2} w_{ij}^{\alpha} d_{ij}. \tag{6}$$

The topological score is designed to capture how disease affects the topological organization of the brain. Its weighted version, which includes the information inherent to the connectivity strength, could in principle enhance the segregation capability of the two cohorts. In fact, we will directly address this aspect in the following sections. The two scores were finally normalized to get a direct probabilistic interpretation as diagnostic scores.

3. Results and Discussion

3.1. Quantitative Assessment of $\mathcal{S}$ and $\mathcal{S}_w$ Scores. To evaluate the capability of both $\mathcal{S}$ and $\mathcal{S}_w$ to capture the effects yielded

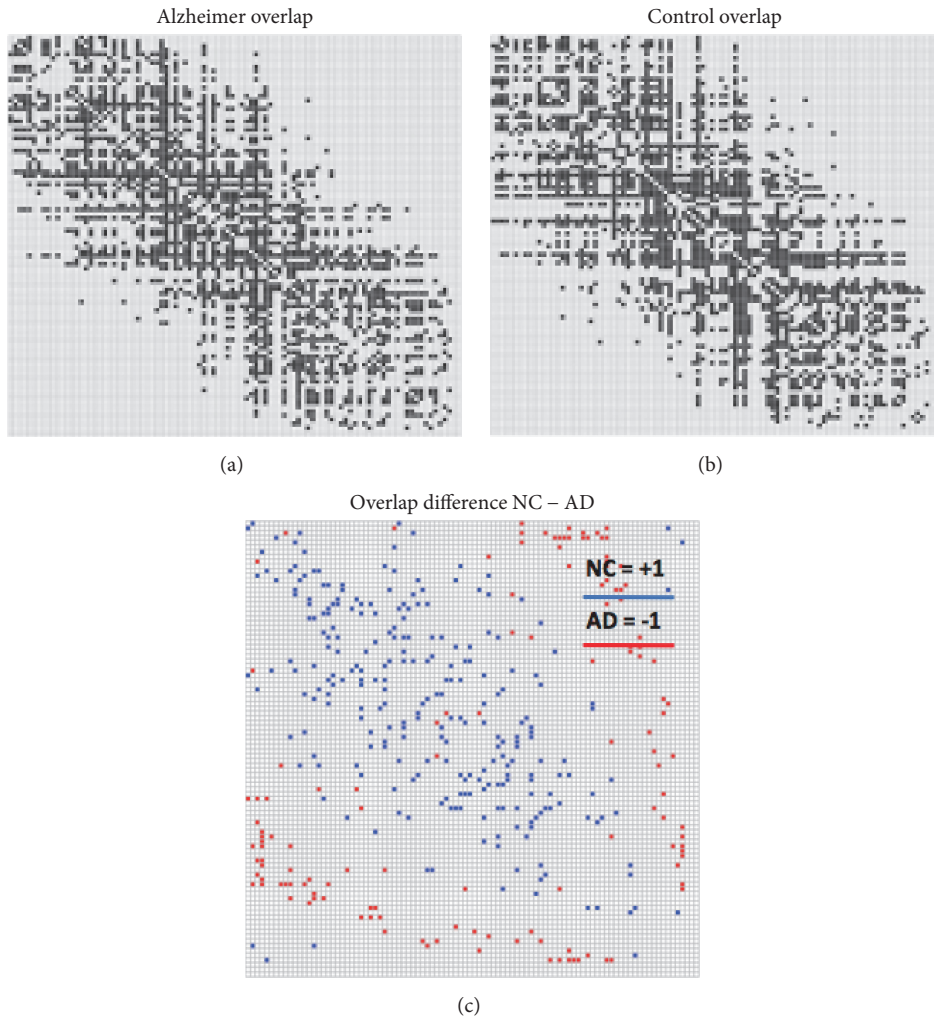


FIGURE 3: The figure shows the overlap matrices computed for the AD (a) and the NC (b) cohorts represented by the angular order of the eigenvectors. (c) shows the overlap difference between the controls and patients. AD patients have a lesser number of edges $E_{AD} = 1463 < 1523 = E_{NC}$. Interestingly, AD and NC seem to show different patterns of connectivity more than an overall impairment.

by disease on brain organization we computed them adopting a leave-one-out cross-validation framework. Thus, each score was computed using the difference overlap $\mathscr{D}$ resulting from the remaining subjects in the cohort. The separation between AD and NC, as shown in Figure 4, denoted a significant effect.

In fact, the topological score $\mathscr{S}$ resulted in a Wilcoxon p value $p = 2 \cdot 10^{-13}$ while for $\mathscr{S}_w$ we found $p_w = 2 \cdot 10^{-11}$. Even if both p values showed a 0.01 significance, the relative effect measured in terms of Cohen's h distance revealed that $\mathscr{S}$ had a larger effect, almost double, than $\mathscr{S}_w$ with $h = 1.4 > 0.8 = h_w$. The effect was also qualitatively manifest when comparing the score distributions, shown in Figures 4(b) and 4(c). The weighted scores of NC and AD showed a greater superimposition if compared with topological scores.

These results demonstrated that the proposed topological scores had a significant association with the disease effect, or in other words, they were proper measurement of the topological differentiation affecting a diseased brain. Provided that the topological score resulted in a diagnostic index being more effective than its weighted variant, they were obviously correlated measures, as shown in Figure 5.

However, their Pearson's correlation $r = 0.61$ was not so high as one could have expected. This result showed that the information carried by both the scores was not redundant. Besides, this result can be interpreted in terms of the quality of the information content carried by both scores. Interestingly, the topological score furnishes better quality information even disregarding the additional weight information. Nonetheless, the weighted topological score should deserve further studies, especially aimed at removing, as previously explained, noisy connections and artifacts yielded by the tractography reconstruction algorithms which obviously negatively affected its discriminating power.

3.2. Brain Topology and Anatomy. Another important aspect concerning the topological score interpretation and its weighted version is whether they can or cannot directly be related to brain anatomy. This analysis in particular aims at

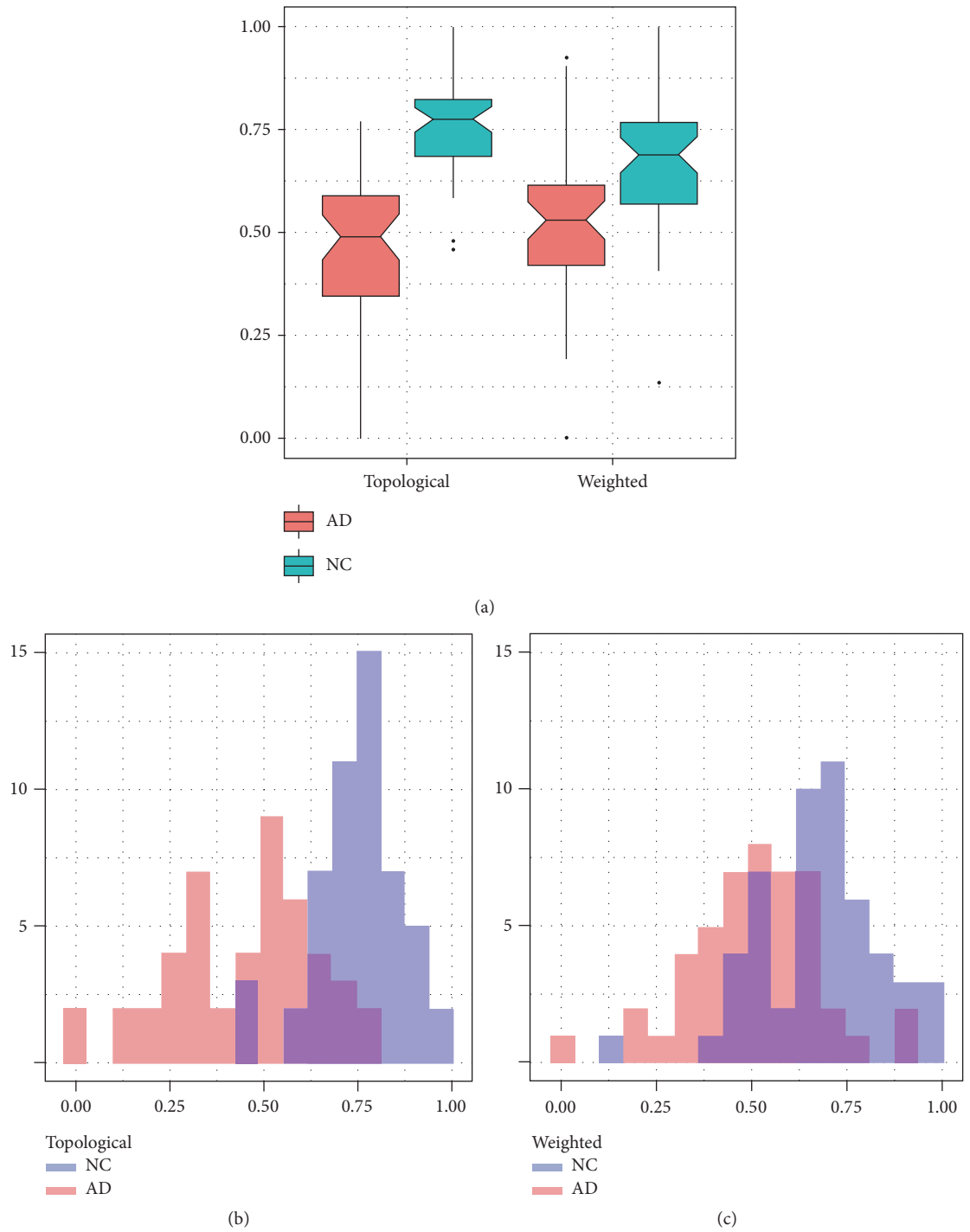


Figure 4: The figure shows (a) the boxplot of topological $\mathcal{S}$ and weighted $\mathcal{S}_w$ scores. The separation effect is more evident when using $\mathcal{S}$. This is also evident when looking at the score distributions: the weighted score (c) shows a consistent overlap between NC and AD if compared with topological score distribution (b).

quantifying whether an association exists from the topological organization of the brain and the atrophy of particular brain regions related to the disease. This test should outline in particular how structural MRI and DWI can be combined to better characterize and distinguish the diseased patterns.

Firstly, we computed the volumes of subcortical features of interest for AD. Specifically, we measured the volumes of Left Thalamus (L-Th), Left-Caudate (L-Cd), Left Putamen (L-Pt), Left Pallidum (L-Pa), Left Hippocampus (L-Hp), Left Amygdala (L-Am), Left Accumbens (L-Ac), Right Thalamus (R-Th), Right Caudate (R-Cd), Right Putamen (R-Pt), Right Pallidum (R-Pa), Right Hippocampus (R-Hp), Right Amygdala (R-Am), and Right Accumbens (R-Ac) with the FSL *FAST* tool. Then we measured Pearson's correlations between

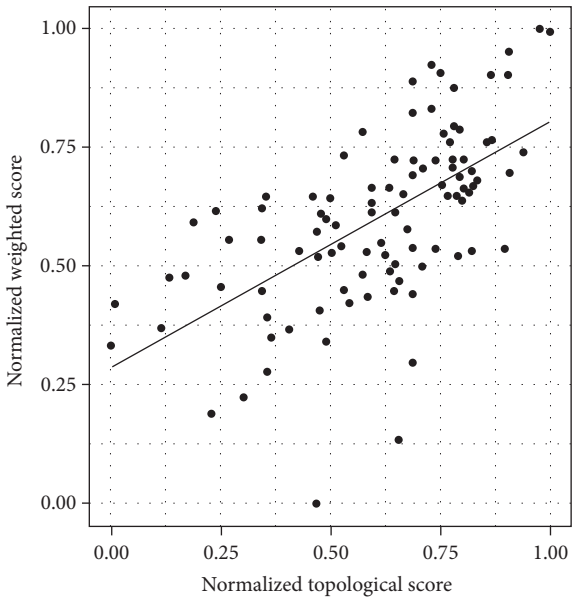


FIGURE 5: The figure shows moderate Pearson's $r = 0.61$ correlation characterizing the topological score $\mathcal{S}$ and its weighted variant $\mathcal{S}_w$. A higher correlation could have been expected; nonetheless, artifacts and noise yielded by reconstruction tractography algorithms have obviously a greater effect on computed weights, more than their presence.

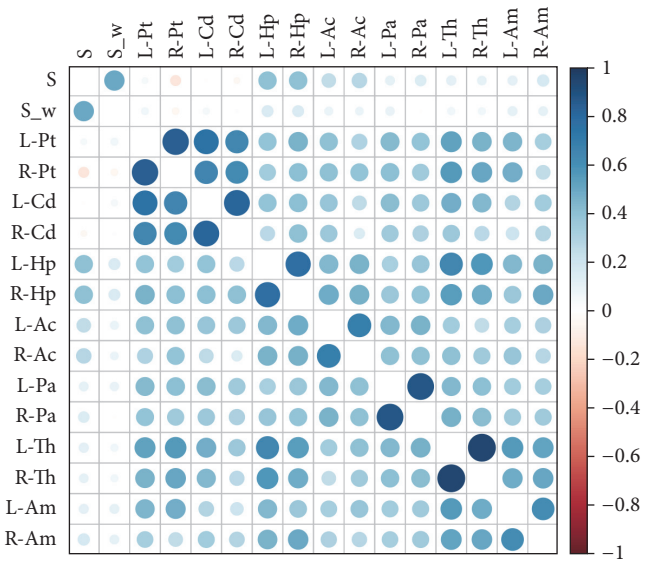


FIGURE 6: The figure shows Pearson's correlation between the proposed topological $\mathcal{S}$ and weighted $\mathcal{S}_w$ scores and the structural measurements of Left Thalamus (L-Th), Left-Caudate (L-Cd), Left Putamen (L-Pt), Left Pallidum (L-Pa), Left Hippocampus (L-Hp), Left Amygdala (L-Am), Left Accumbens (L-Ac), Right Thalamus (R-Th), Right Caudate (R-Cd), Right Putamen (R-Pt), Right Pallidum (R-Pa), Right Hippocampus (R-Hp), Right Amygdala (R-Am), and Right Accumbens (R-Ac).

each regions and our proposed scores $\mathcal{S}$ and $\mathcal{S}_w$. Results are shown in Figure 6.

The correlations were ordered by hierarchical clustering, in this way the more correlated regions tended to be placed together in the correlation matrix. This is the reason, for example, for the manifest pairing of left/right regions. It is worth noting that both $\mathcal{S}$ and $\mathcal{S}_w$ were poorly correlated to structural features. This result would suggest that the topological brain organization contains intrinsic information that it does not share with structural measurements. The most correlated structural features to the proposed scores ($r \sim 0.3$) were the hippocampal volumes.

To measure the information content provided by $\mathcal{S}$ and $\mathcal{S}_w$ we trained with both of them (as we previously demonstrated they were not redundant) a support vector machine model with 500 5-fold cross-validation. Obviously, to avoid any bias in this step the computation of matrices $\mathcal{O}_{\mathrm{NC}}$, $\mathcal{O}_{\mathrm{AD}}$, and $\mathcal{D}$ was performed again, but considering only the training sample. This test allowed also assessing the information contained in $\mathcal{S}$ and $\mathcal{S}_w$, when compared with the structural features derived from T1 scans. For this measure we adopted the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Results are summarized in Figure 7.

The average AUC corresponding to $\mathcal{S}$ and $\mathcal{S}_w$ scores was 95% with a 95% confidence interval of 92%–99%. For what concerns structural features the performance had a drastic drop with an AUC of 76% and confidence interval of 66%–86%. Interestingly, when combining the information of structural features with the topological one not a significant

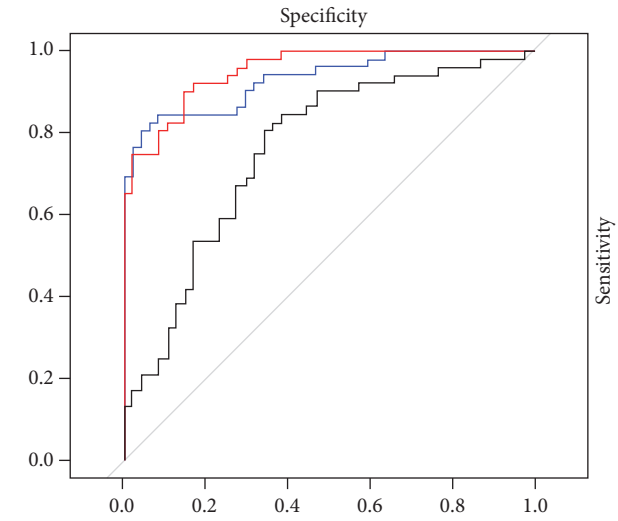


FIGURE 7: A comparison of the receiver operating characteristic curves for $\mathcal{S}$ and $\mathcal{S}_w$ scores (red), the structural features (black), and their combination (blue) is presented. Corresponding AUC performance is 95%, 93%, and 76%.

effect shows up. In fact, AUC was 93% with a 95% confidence interval of 0.88–0.98.

Structural and topological features are not correlated as shown in Figure 6; therefore, one could expect an improvement of classification when combining the two typologies of features. However, as previously mentioned, this is still an open question. For what concerns this study, these results made us hypothesize that there could be a misleading effect

driven by confounding features. To test this hypothesis we used among structural features only hippocampal volumes; the measured AUC (88%, 95% confidence interval 75%–91%) was slightly higher than using the whole set of structural features; its effect also improved, even if not significantly, the overall classification performance with topological features (AUC 97%, 95% confidence interval 94%–100%).

This result suggests that a careful feature selection strategy should be applied to gain an effective information contribution from different imaging modalities.

4. Conclusions

In this study a novel approach to characterize the brain organization from a topological perspective is presented. In particular, because of the well-known and stereotyped pattern characterizing AD, we chose to use this pathology as a benchmark. A topological score and a weighted variant have been defined and used to train support vector machines on a mixed NC/AD cohort. Results showed that topological information was able to efficiently detect diseased patterns (AUC = 95%, 95% confidence interval 92%–99%).

We also addressed in this study the problem of quantifying the effect of combining MRI-based features with topological ones. We found that their combination can improve classification accuracy; nonetheless, this is strictly related to the quality of structural features used. In fact, when using all MRI features available the classification performance decreased; on the contrary, it was slightly raised using hippocampal volumes whose association with AD is well known. A subtle effect should be better investigated on larger cohorts.

The performance obtained is comparable with best results reported in the literature so far, but possible improvements could include a more refined study of weighted networks, instead of their binary version; nevertheless, this cannot be considered a limitation of the present study, whose main goal was to investigate the brain topology and understand whether the topological measures proposed were suitable for clinical purposes.

The presented methodology is general, even if in this case it has been tailored on Alzheimer's disease. For future work, we propose to investigate the application of this methodology to mixed cohorts including also MCI subjects, trying to tackle the discrimination problem between subjects converting to AD or not, and the early diagnosis of AD. Patients affected by neurodegenerative diseases incur a cognitive impairment which could be effectively diagnosed and monitored by these measurements, a useful trait for technological innovations in the e-health field, for example, for remote medicine applications, or for pharmacological industries, aiming at the development of drug therapies and clinical trials. Further investigations could be aimed at diseases affecting the brain organization with less stereotyped patterns.

Disclosure

Data used in preparation of this article were obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Competing Interests

All authors disclose no actual or potential conflict of interests, including any financial, personal, or other relationships with other people or organizations that could inappropriately influence their work.

Acknowledgments

N. Amoroso acknowledges funding by the Italian MIUR Grant PON PRISMA Cod. PON04a2_A. This research was also supported by Istituto Nazionale di Fisica Nucleare (INFN), Italy. Data collection and sharing for this project were funded by Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense Award no. W81X WH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Homann-La Roche Ltd. and its aliated company Genentech, Inc.; GE Healthcare; Innogenetics, NV; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institute of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] G. B. Frisoni, N. C. Fox, C. R. Jack Jr., P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nature Reviews Neurology*, vol. 6, no. 2, pp. 67–77, 2010.

[2] G. Nicoletti, R. Lodi, F. Condino et al., "Apparent diffusion coefficient measurements of the middle cerebellar peduncle differentiate the Parkinson variant of MSA from Parkinson's disease and progressive supranuclear palsy," *Brain*, vol. 129, no. 10, pp. 2679–2687, 2006.

[3] A. Chincarini, P. Bosco, G. Gemme et al., "Alzheimer's disease markers from structural MRI and FDG-PET brain images," *European Physical Journal Plus*, vol. 127, no. 11, article no. 135, 2012.

[4] N. Amoroso, R. Errico, G. Ferraro, S. Tangaro, A. Tateo, and R. Bellotti, "Fully automated MRI analysis for brain diseases with high performance computing," in *Proceedings of the SCORE@POLIBA Workshop*, pp. 347–351, December 2014.

[5] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[6] R. Bellotti, A. Lombardi, N. Amoroso, A. Tateo, and S. Tangaro, "Semi-unsupervised prediction for mild TBI based on Both Graph and K-nn methods," in *Proceedings of the International Workshop on Brainlesion: mTOP2016*, Springer, 2016.

[7] D. E. Barnes and K. Yaffe, "The projected effect of risk factor reduction on Alzheimer's disease prevalence," *The Lancet Neurology*, vol. 10, no. 9, pp. 819–828, 2011.

[8] B. Dubois, H. H. Feldman, C. Jacova et al., "Revising the definition of Alzheimer's disease: a new lexicon," *The Lancet Neurology*, vol. 9, no. 11, pp. 1118–1127, 2010.

[9] G. I. Allen, N. Amoroso, C. Anghel et al., "Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease," *Alzheimer's & Dementia*, vol. 12, no. 6, pp. 645–653, 2016.

[10] E. E. Bron, M. Smits, W. M. van der Flier et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge," *NeuroImage*, vol. 111, pp. 562–579, 2015.

[11] N. Amoroso, R. Errico, and R. Bellotti, "PRISMA-CAD: fully automated method for computer-aided diagnosis of dementia based on structural MRI data," in *Proceedings of the MICCAI workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*, pp. 16–23, September 2014.

[12] N. Amoroso, R. Errico, S. Bruno et al., "Hippocampal unified multi-atlas network (HUMAN): protocol and scale validation of a novel segmentation tool," *Physics in Medicine and Biology*, vol. 60, no. 22, article 8851, 2015.

[13] A. Chincarini, F. Sensi, L. Rei et al., "Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease," *NeuroImage*, vol. 125, pp. 834–847, 2016.

[14] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[15] C.-Y. Wee, P.-T. Yap, D. Zhang et al., "Identification of MCI individuals using structural and functional connectivity networks," *NeuroImage*, vol. 59, no. 3, pp. 2045–2056, 2012.

[16] M. Dyrba, M. Grothe, T. Kirste, and S. J. Teipel, "Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM," *Human Brain Mapping*, vol. 36, no. 6, pp. 2118–2131, 2015.

[17] G. T. Stebbins and C. M. Murphy, "Diffusion tensor imaging in Alzheimer's disease and mild cognitive impairment," *Behavioural Neurology*, vol. 21, no. 1-2, pp. 39–49, 2009.

[18] S. Kitamura, K. Kiuchi, T. Taoka et al., "Longitudinal white matter changes in Alzheimer's disease: a tractography-based analysis study," *Brain Research*, vol. 1515, pp. 12–18, 2013.

[19] T. M. Nir, N. Jahanshad, J. E. Villalon-Reina et al., "Effectiveness of regional DTI measures in distinguishing Alzheimer's disease, MCI, and normal aging," *NeuroImage: Clinical*, vol. 3, pp. 180–195, 2013.

[20] H. Huang, X. Fan, M. Weiner et al., "Distinctive disruption patterns of white matter tracts in Alzheimer's disease with full diffusion tensor characterization," *Neurobiology of Aging*, vol. 33, no. 9, pp. 2029–2045, 2012.

[21] A. Fellgiebel, P. Wille, M. J. Müller et al., "Ultrastructural hippocampal and white matter alterations in mild cognitive impairment: a diffusion tensor imaging study," *Dementia and Geriatric Cognitive Disorders*, vol. 18, no. 1, pp. 101–108, 2004.

[22] J. Acosta-Cabronero, G. B. Williams, G. Pengas, and P. J. Nestor, "Absolute diffusivities define the landscape of white matter degeneration in Alzheimer's disease," *Brain*, vol. 133, no. 2, pp. 529–539, 2010.

[23] L. Zhan, J. Zhou, Y. Wang et al., "Comparison of 9 tractography algorithms for detecting abnormal structural brain networks in Alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 7, article 48, 2015.

[24] M. Bastiani, N. J. Shah, R. Goebel, and A. Roebroeck, "Human cortical connectome reconstruction from diffusion weighted MRI: the effect of tractography algorithm," *NeuroImage*, vol. 62, no. 3, pp. 1732–1749, 2012.

[25] G. Prasad, S. H. Joshi, T. M. Nir, A. W. Toga, and P. M. Thompson, "Brain connectivity and novel network measures for Alzheimer's disease classification," *Neurobiology of Aging*, vol. 36, S1, pp. S121–S131, 2015.

[26] A. D. Friederici, J. Bahlmann, S. Heim, R. I. Schubotz, and A. Anwander, "The brain differentiates human and non-human grammars: functional localization and structural connectivity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 7, pp. 2458–2463, 2006.

[27] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[28] O. Sporns, G. Tononi, and G. M. Edelman, "Theoretical neuroanatomy and the connectivity of the cerebral cortex," *Behavioural Brain Research*, vol. 135, no. 1-2, pp. 69–74, 2002.

[29] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, "Detecting rich-club ordering in complex networks," *Nature Physics*, vol. 2, no. 2, pp. 110–115, 2006.

[30] M. P. van den Heuvel and O. Sporns, "Rich-club organization of the human connectome," *The Journal of Neuroscience*, vol. 31, no. 44, pp. 15775–15786, 2011.

[31] R. Migliaccio, F. Agosta, K. L. Possin, G. D. Rabinovici, B. L. Miller, and M. L. Gorno-Tempini, "White matter atrophy in Alzheimer's disease variants," *Alzheimer's and Dementia*, vol. 8, no. 5, pp. S78–S87, 2012.

[32] C.-Y. Lo, P.-N. Wang, K.-H. Chou, J. Wang, Y. He, and C.-P. Lin, "Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in Alzheimer's disease," *Journal of Neuroscience*, vol. 30, no. 50, pp. 16876–16885, 2010.

[33] T. M. Nir, N. Jahanshad, A. W. Toga et al., "Connectivity network measures predict volumetric atrophy in mild cognitive impairment," *Neurobiology of Aging*, vol. 36, no. 1, pp. S113–S120, 2015.

[34] J. A. Brown, K. H. Terashima, A. C. Burggren et al., "Brain network local interconnectivity loss in aging APOE-4 allele carriers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 51, pp. 20760–20765, 2011.

[35] M. Daianu, N. Jahanshad, T. M. Nir et al., "Breakdown of brain connectivity between normal aging and Alzheimer's disease: a structural k-Core network analysis," *Brain Connectivity*, vol. 3, no. 4, pp. 407–422, 2013.

[36] B. M. Tijms, A. M. Wink, W. de Haan et al., "Alzheimer's disease: connecting findings from graph theoretical studies of brain networks," *Neurobiology of Aging*, vol. 34, no. 8, pp. 2023–2036, 2013.

[37] M. Daianu, N. Jahanshad, T. M. Nir et al., "Rich club analysis in the Alzheimer's disease connectome reveals a relatively undisturbed structural core network," *Human Brain Mapping*, vol. 36, no. 8, pp. 3087–3103, 2015.

[38] G. Pergola, S. Trizio, P. Di Carlo et al., "Grey matter volume patterns in thalamic nuclei are associated with familial risk for schizophrenia," *Schizophrenia Research*, vol. 180, pp. 13–20, 2017.

[39] S. Boccaletti, G. Bianconi, R. Criado et al., "The structure and dynamics of multilayer networks," *Physics Reports. A Review Section of Physics Letters*, vol. 544, no. 1, pp. 1–122, 2014.

[40] G. Bianconi, "Statistical mechanics of multiplex networks: entropy and overlap," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 87, no. 6, Article ID 062806, 2013.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 9818506, 10 pages
https://doi.org/10.1155/2017/9818506

Research Article

3D Kidney Segmentation from Abdominal Images Using Spatial-Appearance Models

Fahmi Khalifa,[1,2] Ahmed Soliman,[1] Adel Elmaghraby,[3] Georgy Gimel'farb,[4] and Ayman El-Baz[1]

[1]*Bioengineering Department, University of Louisville, Louisville, KY, USA*
[2]*Electronics and Communication Engineering Department, Mansoura University, Mansoura, Egypt*
[3]*Computer Engineering and Computer Science Department, University of Louisville, Louisville, KY, USA*
[4]*Department of Computer Science, University of Auckland, Auckland, New Zealand*

Correspondence should be addressed to Ayman El-Baz; aselba01@louisville.edu

Received 19 August 2016; Revised 29 November 2016; Accepted 22 December 2016; Published 9 February 2017

Academic Editor: Po-Hsiang Tsui

Kidney segmentation is an essential step in developing any noninvasive computer-assisted diagnostic system for renal function assessment. This paper introduces an automated framework for 3D kidney segmentation from dynamic computed tomography (CT) images that integrates discriminative features from the current and prior CT appearances into a random forest classification approach. To account for CT images' inhomogeneities, we employ discriminate features that are extracted from a higher-order spatial model and an adaptive shape model in addition to the first-order CT appearance. To model the interactions between CT data voxels, we employed a higher-order spatial model, which adds the triple and quad clique families to the traditional pairwise clique family. The kidney shape prior model is built using a set of training CT data and is updated during segmentation using not only region labels but also voxels' appearances in neighboring spatial voxel locations. Our framework performance has been evaluated on in vivo dynamic CT data collected from 20 subjects and comprises multiple 3D scans acquired before and after contrast medium administration. Quantitative evaluation between manually and automatically segmented kidney contours using Dice similarity, percentage volume differences, and 95th-percentile bidirectional Hausdorff distances confirms the high accuracy of our approach.

1. Introduction

Kidney segmentation from dynamic contrast-enhanced computed tomography (CT) is of immense importance for any computer-assisted diagnosis of renal function assessment, pathological tissue localization, radiotherapy planning, and so forth [1]. Nevertheless, accurate segmentation of kidney tissues from dynamic CT images is challenging due to many reasons, including data acquisition artifacts, large inhomogeneity of the kidney (e.g., cortex and medulla), large anatomical differences between subjects, similar intensities of adjacent organs, and varying signal intensities over the time course of data collection due to agent transit [2, 3]; see Figure 1.

Many automated and semiautomated approaches have been developed to address these challenges. Earlier computerized renal image analysis (e.g., [4]) was usually carried out either manually or semiautomatically. Typically, a user-defined region-of-interest (ROI) is delineated in one image and for the rest of the images, image edges were detected and the model curve was matched to these edges. However, ROI placements are based on the users' knowledge of anatomy and thus are subject to inter- and intraobserver variability. Additionally, these methods are very slow, even though semiautomated techniques reduce the processing time. Traditional segmentation techniques utilizing image thresholding or region growing [5–9] have been also explored for CT kidney segmentation. For example, Pohle and Toennies [7] developed an automatic region-growing algorithm based on estimating the homogeneity criterion from the characteristics of the input images. A semiautomated method was also proposed by Leonardi et al. [9]. First, a region-growing approach

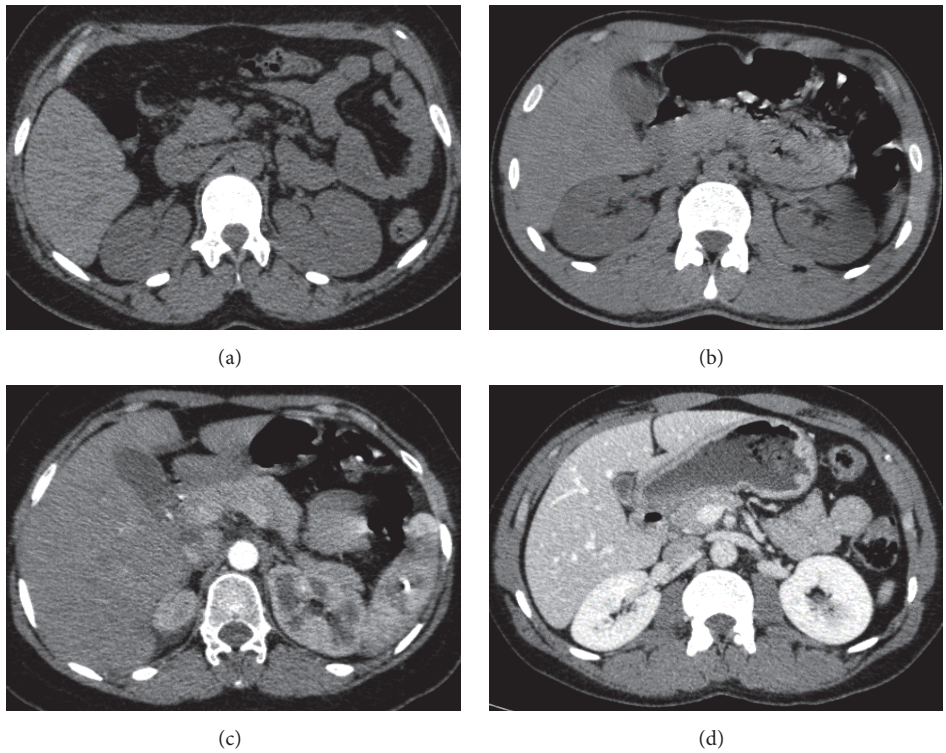


Figure 1: Axial cross-sectional images showing different CT data challenges: (a) low contrast, (b) intensity inhomogeneities, (c) fuzzy boundary, and (d) contrast and anatomy differences.

is performed to obtain an initial kidney segmentation from the grayscale image stack. Then, a refinement step utilizing histogram analysis is employed to the initially segmented kidney regions to reduce incorrectly segmented areas. However, these traditional methods are not accurate due to the large overlap of gray level intensity between the kidney and other surrounding tissues in addition to sensitive to initialization.

To more accurately segment abdominal CTs, recent segmentation methods consider either visual appearances, prior shapes, texture features, or hybrid techniques. In particular, Tsagaan et al. [10] presented a deformable model-based approach utilizing a nonuniform rational B-spline surface representation. Their framework incorporated statistical shape information (e.g., mean and variation) into the objective function for the model fitting process as an additional energy term.

A constrained optimization deformable contour by Wang et al. [11] exploited the degree of contour interior homogeneity as an extra constraint within the level set energy minimization framework. Lu et al. [12] developed a coarse-to-fine approach for kidney segmentation on abdominal CT images using the Chan-Vese (CV) level set method [13]. Mathematical morphology operations are performed to extract the kidney structures interactively with prior anatomy knowledge. Huang et al. [14] proposed a multiphase level set approach with multidynamic shape models to segment the kidneys on abdominal CT images. Campadelli et al. [15] proposed an automatic, gray-level based segmentation framework based on a multiplanar fast marching method. A stochastic level set-based framework by Khalifa et al. [16, 17] integrated probabilistic kidney shapes and image signals priors into Markov random field (MRF) for abdominal 3D CT kidney segmentation. Despite their popularity, deformable model-based methods fail in the case of excessive image noise, poor image resolution, or diffused boundaries if they do not take advantage of a priori models.

Freiman et al. [18] proposed a model-based framework utilizing maximum a posteriori-MRF (MAP-MRF) estimation of the input CT image. The MAP-MRF estimation is obtained by using a graph min–cut technique. Lin et al. [19] proposed a framework that combined region- and model-based methods. Initial kidney location is estimated using geometrical location, statistical information, and a priori anatomical knowledge. Secondly, an elliptic candidate kidney region extraction approach is proposed. Finally, an adaptive region-growing approach is employed for kidney segmentation. Spiegel et al. [20] proposed an active shape model (ASM) based framework that was combined with a curvature-based nonrigid registration approach to solve the point correspondence problem of the training data. A hybrid framework by Chen et al. [21] combined active appearance model (AAM), live wire, and graph-cuts methods for 3D abdominal organ segmentation. In general, parametric shape-based techniques depend on the existence of adequate texture features in abdominal images and may perform poorly due to noise and the lack of well-defined features. Cuingnet et al. [22] exploited random regression and classification forests for CT kidney images segmentation. Initially, global contextual

information is used to detect the kidney. This is followed by a cascade of local regression forests for refinement. Then, probabilistic segmentation maps are built using classification forest. Finally, an implicit template deformation algorithm driven by these maps is employed to obtain the final segmentation. A model-based framework by Badakhshannoory and Saeedi [23] combined low-level segmentation schemes with a statistical-based modeling approach. First, an organ space is built using a statistical model and principle component analysis. Then, each image slice of an input CT volume is segmented multiple times using a graph-based segmentation by varying segmentation parameters. Finally, a distance-based criterion from the organ space is used to choose the closest candidate as the best segmentation result. In general, knowledge-based approaches are computationally intensive, and their accuracy depends on the training data size.

Bagci et al. [24] developed a multiobject segmentation framework that integrates a statistical shape model and hierarchical object recognition into a global graph-cuts segmentation model. Wolz et al. [25] developed a hierarchical two-step atlas registration framework for multiobject segmentation. First, subject-specific priors are generated from an atlas database based on multiatlas registration and patch-based segmentation. Final segmentation is obtained using graph-cuts, incorporating high-level spatial knowledge and a learned intensity model. Another study by Okada et al. [26] performed multiobject segmentation using probabilistic atlases that combines interorgan spatial and intensity a priori models. Despite the potential to improve the segmentation accuracy due to the spatial kidney constraints from other organs, multiobject segmentation schemes require more comprehensive prior information. A semiautomated Grow-Cut algorithm by Dai et al. [27] employed a monotonically decreasing function and image gray features to propagate initial user-defined labels over all the slices to derive an optimal cut for a given CT data in space. Zhao et al. [28] proposed a sliced-based framework for 3D kidney segmentation. First, an initial segmentation is obtained using the CV approach [13]. Then, a set of contextual features (e.g., slices overlap, the distance) and multiple morphological operations are used to estimate the continuity between slices. The final segmentation is obtained by discarding the leakage and the weak edges between adjacent slices using a local iterative thresholding method. Chu et al. [29] presented an automated MAP-based multiorgan segmentation method that incorporated image-space division and multiscale weighting scheme. Their framework is based on a spatially divided probabilistic atlases and the segmentation is performed using a graph cut method. Yang et al. [30] developed on multiatlas framework using a two-step approach to obtain coarse-to-fine kidney segmentation. A coarse segmentation is obtained by registering an input down-sampled CT volume with a set of low-resolution atlas images. Then, cropped kidney images are coaligned with high-resolution atlas images using B-Splines registration. The final segmentation result is obtained by majority voting of all deformed labels of all atlas images. Liu et al. [31, 32] developed a framework for kidney segmentation on noncontrast CT images using efficient belief propagation. A preprocessing step is applied to extract anatomical landmarks to localize kidney search regions. Then, an efficient belief propagation is used to extract the kidney by minimizing an energy function that incorporates intensity and prior shape information. However, the method was evaluated on five noncontrast CT data sets only and additional segmentation of other organs (e.g., liver, spleen) is required to determine subimages that envelope the kidneys.

In summary, during the last few years there have been numerous studies for abdominal CTs kidney segmentation. In addition to the above-mentioned limitations, current methods have the following shortcomings. Most of them are based on visual appearance and did not take into account the spatial interaction relationships. Most of the shape-based methods utilize fixed models and therefore have limited accuracy for CT data outside their training scope. Most of the existing methods work very well with contrast CTs only. Most of the energy-based methods (e.g., graph-cut) use regional and boundary information that may not exist in some (e.g., precontrast) images and may not achieve globally optimal results.

To account for these limitations, we developed a 3D kidney segmentation framework that integrates, in addition to the current CT appearance features, higher-order appearance models and adaptive shape model features into a random forests (RF) classification model [33]. The integrated features increase the ability of our framework to account for the large CT images' inhomogeneities and therefore accurately segment both contrast and noncontrast CTs. Particularly, the spatial features are based on a higher-order Markov–Gibbs random field (MGRF) model that adds to the traditional pairwise cliques [34] the families of the triple and quad cliques. The spatial-appearance kidney shape prior is an adaptive model that is updated during segmentation and accounts not only for region labels, but also intensities in neighboring spatial locations. Moreover, compared to other tissue classification methods the RF is employed due to its (i) powerful ability to learn the characteristics of complex data classes [35], (ii) less sensitivity to data outliers, (iii) ability to overcomes overfitting of the training set, and (iv) ability to handle high dimensional spaces as well as large number of training examples.

A detailed description of our developed methodology for kidney segmentation from dynamic CT images including the details of the discriminative features is given in Section 2. It is worth mentioning that, in addition to our methodology presentation in [33], this paper provides (i) a more comprehensive review of the related literature work on the abdominal CT images segmentation (Section 3); (ii) detailed description of the metrics that are used for segmentation evaluation of our and compared techniques (Section 3); and (iii) expansion of the experimental results by adding an essential metric that is used to evaluate the robustness of segmentation techniques, namely, the receiver operating characteristics (ROC) (Section 4).

2. Methods

A block diagram of our kidney segmentation framework is shown in Figure 2. Our technique is based on random forests

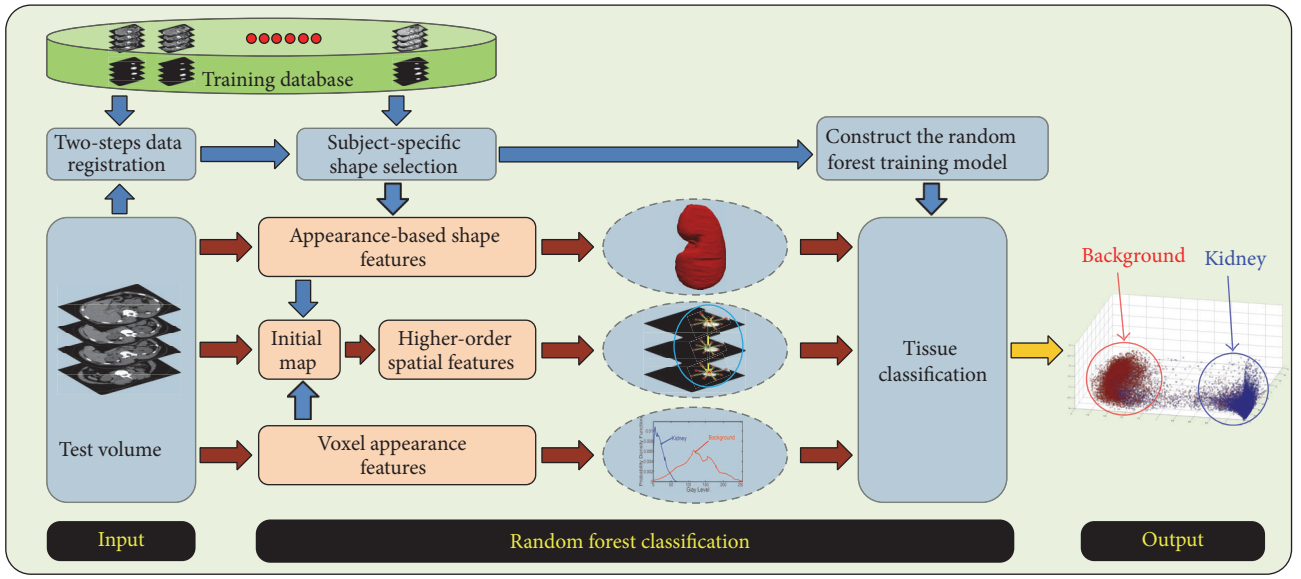


Figure 2: Block diagram of our kidney segmentation framework from abdominal CT images using random forest (RF).

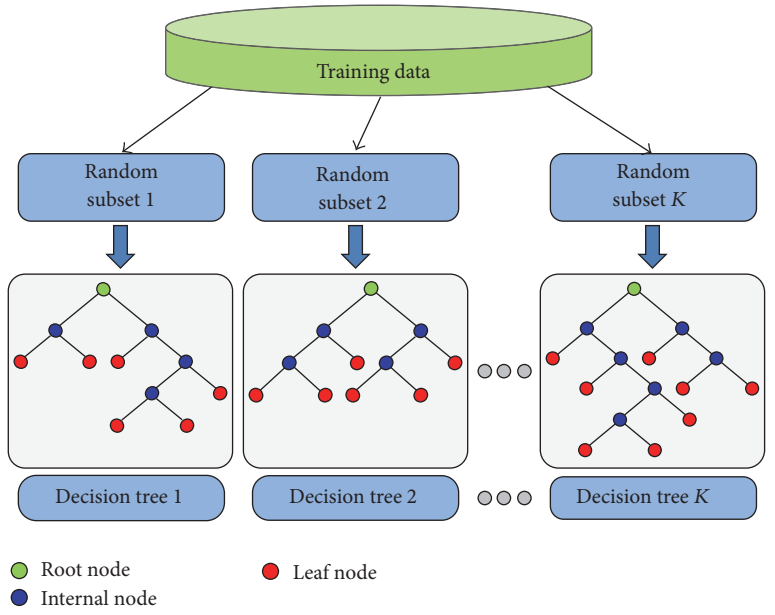


Figure 3: A schematic illustration of the random decision trees for random forests (RF) classification.

(RF) classification and incorporates spatial-appearance features for better separation of the CT data classes. RF is an efficient multiclass machine learning technique, which is increasingly being utilized in data clustering as well as image classification. As an ensemble learning classifier, RF typically consists of many decision trees (DTs) and combines two main concepts [36]. The first is the random selection of features and the second is "bagging" [37], which implies the training of each DT with a randomly chosen and overlapping subset of the training samples. In general, as numbers of the DTs increase the results get better. Nevertheless, there is a threshold beyond which the performance benefit from adding more DTs will be lower than the computational cost for learning these additional DTs [38].

During the RF training phase, each DT recursively processes its randomly selected training samples' features along a path starting from the tree's root node using binary classification tests, as shown in Figure 3. The latter tests compare the features' values at each internal tree node to a certain threshold that is selected using a certain criterion. A leaf node of the DT is reached if all samples belong to a single class; the number of data samples is smaller than a predefined value, or the maximum tree depth is reached [35]. Once occurred, the most frequent class label of the training

data at the node is stored for the testing phase. For testing, a given data sample is handled by applying respective tests in line with the path it traverses from the tree root node to the leaf. When a leaf node is reached, the DT casts a vote corresponding to the class assigned to this node in the training stage. Finally, a majority voting is used to class-label test samples. The final class probabilities are estimated by the fraction of votes for that class by all DTs.

In order to build an accurate RF model that provides better separation of data classes, discriminative and robust features are needed. Therefore, in this paper multiple features from the CT data are extracted, for both training and testing phases. These features include (i) first-order appearance (Hounsfield units (Hus) values) features; (ii) higher-order spatial interaction features; and (iii) appearance-based shape model features. Those features are extracted at each voxel's location $\mathbf{p} = (x, y, z)$ in the 3D arithmetic lattice $\mathbf{R} = \{(x, y, z)\ 0 \leq x \leq X-1,\ 0 \leq y \leq Y-1,\ 0 \leq z \leq Z-1\}$ supporting the grayscale CT images, $\mathbf{g} = \{g_{\mathbf{p}} : \mathbf{p} \in \mathbf{R},\ g_{\mathbf{p}} \in \mathbf{Q}\}$, and their region, or segmentation maps, $\mathbf{m} = \{m_{\mathbf{p}} : \mathbf{p} \in \mathbf{R}, m_{\mathbf{p}} \in \mathbf{L}\}$. Here, $\mathbf{Q} = \{0, 1, \ldots, Q-1\}$ and $\mathbf{L} = \{\text{“KT”}, \text{“OT”}\}$ is a finite set of integer gray levels and region labels (kidney object tissues “KT” and other background tissues “OT”), respectively. Since spatial and shape features are based on probabilistic models, the first-order appearance-based features were also normalized to reduce the domination of a specific feature during RF classification. Details of the employed features are given in the following sections.

2.1. First-Order Appearance Features. The first type of features that are used in our framework is the CT voxel-appearance features. Those features were extracted at each voxel $\mathbf{p}$ regionally from the CT data after normalization. Due to image noise presence and reconstruction artifacts, we used, at each voxel $\mathbf{p}$, regional intensity features in addition to the local CT Hounsfield units (HU). Namely, we used the mean HU values of a symmetric 3D cube (i.e., voxels' 26-neighbors) centered around $\mathbf{p}$ and the mean of the HUs of a 3×3 in-plane symmetric window (i.e., voxels' 8-neighbors) centered around $\mathbf{p}$.

2.2. Shape Prior Features. The ultimate goal is to accurately segment the kidney from the CT data such that the extracted kidney borders closely approximate the expert manual delineation. However, due to the similar visual appearance between some kidney structures (e.g., medulla) and background, the segmentation should not rely only on image signals. Therefore, shape features of the expected kidney shape are also used in our segmentation framework. In this paper, we employed an adaptive, probabilistic kidney shape model that takes into account not only voxels' location, but also their intensity information [39, 40].

For training, a shape database is constructed using a set of training data sets that is collected from different subjects; each contains multiple CT scans acquired at different phases of contrast-enhancements. The ground truth segmentation (labeled data) of the training images is obtained by manual delineation of the kidney borders by an expert. In order to reduce the variability across subjects and maximize overlaps of the kidneys for estimating the shape prior probability, the training grayscale images are coaligned using a two-step registration methodology. First, a 3D affine transformation is used with 12 degrees of freedom (3 for the 3D translation, 3 for the 3D rotation, 3 for the 3D scaling, and 3 for the 3D shearing) to account for global motion [41]. Second, local kidney deformations are handled using a 3D B-splines based transformation proposed in [42]. Finally, the obtained transformation parameters for each scan are applied to its binary (labeled) data to be used during segmentation to estimate the shape prior probability.

For testing, an input grayscale 3D CT kidney image, $\mathbf{g_t}$, to be segmented is first coaligned with the training database using the two-step registration methodology described above. Then, a subject-specific shape, $\mathbf{g_i}$, $i = 1, 2, \ldots, N$, is extracted by computing the conventional normalized cross correlations (NCC) between the coaligned input grayscale image and all grayscale images in the database, to select the top N similar kidneys ($N = 19$ in our experiments below). Finally, visual appearances of both the input 3D grayscale CT image and the selected grayscale training images guide adapting the shape prior. Namely, the voxel-wise probabilities, $P_{s:\mathbf{p}}(l)$ for the adaptive shape prior $P_s(\mathbf{m}) = \prod_{\mathbf{p}\in\mathbf{R}} P_{s:\mathbf{p}}(m_p)$, are estimated based on the found voxels $l \in \mathbf{L}$. Let $\mathbf{v}_{i:\mathbf{p}}(l) = \{\rho : \rho \in \mathbf{R}; \rho \in \mathbf{C_p}; |g_{i:\rho} - g_{t:\mathbf{p}}| \leq \tau\}$ be a subset of similar training voxels within a search cube $\mathbf{C_p}$ in the training image g_i, where τ is a predefined fixed signal range and $g_{t:p}$ is the mapped input signal. Let $v_{i:\mathbf{p}} = \text{card}(\mathbf{v}_{i:\mathbf{p}})$ denote the cardinality (number of voxels) of this subset $\mathbf{v_p} = \sum_{i=1}^{N} \mathbf{v}_{i:\mathbf{p}}$ and $\delta(z)$ be the Kronecker's delta-function: $\delta(0) = 1$ and 0 otherwise. Then $P_{s:\mathbf{p}}(l)$ is given as [39]

$$P_{s:\mathbf{p}}(l) = \frac{\mathbf{1}}{v_{\mathbf{p}}} \sum_{i=1}^{N} \sum_{\boldsymbol{\rho}\in\mathbf{v}_{i:\mathbf{p}}} \boldsymbol{\delta}\left(\mathbf{l} - m_{i:\boldsymbol{\rho}}\right). \tag{1}$$

More details about the adaptive shape model can be found in [39, 40]. Our experiments were conducted using three shape features, like the voxel-appearance features. Namely, we used the $P_s(\mathbf{m})$ value at $\mathbf{p}$, the average $P_s(\mathbf{m})$ value for the 26 neighbors of a 3D cube around $\mathbf{p}$, and the average $P_s(\mathbf{m})$ of the 8 in-plane neighbors for a 3×3 symmetric window centered at $\mathbf{p}$.

2.3. Spatial Features. To improve the segmentation accuracy and account for the large inhomogeneity of the kidney, we incorporated into our segmentation approach the spatial features that describe the relationships between the kidney voxels and their neighbors. These relationships are described using a higher-order spatial model with analytically estimated potentials. The spatial modeling enhances the segmentation by calculating the likelihood of each voxel to be kidney or background on the basis of the initial labeling, $\mathbf{m}$, of the adjacent voxels, formed by a voxel-wise classification using shape and intensity values Our spatial interactions model adds the triple and quad clique families to the traditional pairwise clique family [34] using the 18-connectivity neighborhood. Thus, it is an extension of the conventional Potts model [43],

differing only in that the potentials are estimated analytically. For more mathematical details about our higher-order spatial model, please see [33, 44]. Similar to the other features, three spatial-based features were used: the local spatial probability at $\mathbf{p}$ and the average probabilities for a 3D cube and a 3×3 window centered around $\mathbf{p}$. In total, the whole segmentation approach is summarized in Algorithm 1.

Algorithm 1 (3D kidney segmentation steps).

Step 1 (data coalignment and shape database selection)

(a) Register the input grayscale CT volume to the training database using the two-step registration in Section 2.2.

(b) Calculate the NCC between the input coaligned data and all training volumes. Then, select the NCC-19-top ranked training samples.

Step 2 (features extraction)

(a) Estimate the voxel-appearance features of the coaligned CT volume.

(b) Estimate the higher-order Potts-MGRF spatial probabilities $P_G(\mathbf{m})$.

(c) Estimate the appearance-based shape prior $P_s(\mathbf{m})$ using the method described in [39, 40].

Step 3 (RF training)

(a) Construct the RF training model for the selected 19-top-ranked training images.

Step 4 (tissue segmentation)

(a) Obtain the final segmentation of the input CT volume using the model in Step 3.

3. Segmentation Evaluation Metrics

The performance of our segmentation is evaluated using two metrics. The first is a volumetric-based similarity that characterizes spatial overlaps and volume differences between the segmented and "ground-truth" kidney regions. This type of metrics is important for studying area measurements, for example, total kidney volumes. The second is a distance-based metric that measures how close the edge of a segmented region is to the ground truth, that is, how accurate the shape of a segmented object is with respect to ground truth. Here, we used the Dice coefficient (DC) and percentage volume difference (PVD) to describe the volumetric-based similarity, while the bidirectional 95th-percentile Hausdorff distance (BHD_{95}) is used to characterize the distance-based error metric: $\mathbf{G} \leftrightarrow \mathbf{S}$.

Let $\mathbf{G}$ and $\mathbf{S}$ denote sets of ground-truth and segmented kidney voxels, respectively. The similarity volumetrics evaluate an overlap between these sets and account for cardinalities (i.e., voxel numbers) $c_i = |V_i|$ of true positive (tp), false positive (fp), and false negative (fn) subsets V_i; $i \in \{\mathrm{tp}, \mathrm{fp}, \mathrm{fn}\}$; see Figure 4.

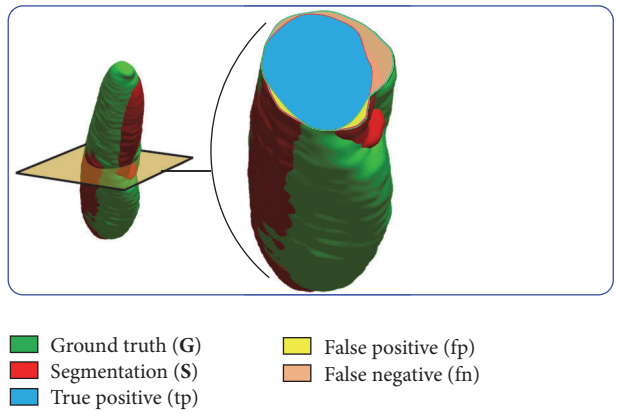


Figure 4: 3D illustration of DC measurement for segmentation evaluation between the ground truth $\mathbf{G}$ and model segmentation $\mathbf{S}$.

The subsets contain true kidney voxels labeled as kidney, nonkidney (background) voxels labeled as kidney, and true kidney voxels labeled as background, respectively:

$$\begin{aligned} V_{\mathrm{tp}} &= \{v : v \in \mathbf{G}, v \in \mathbf{S}\}; \quad c_{\mathrm{tp}} = \left|V_{\mathrm{tp}}\right| \\ V_{\mathrm{fp}} &= \{v : v \notin \mathbf{G}, v \in \mathbf{S}\}; \quad c_{\mathrm{fp}} = \left|V_{\mathrm{fp}}\right| \\ V_{\mathrm{fn}} &= \{v : v \in \mathbf{G}, v \notin \mathbf{S}\}; \quad c_{\mathrm{fn}} = \left|V_{\mathrm{fn}}\right|. \end{aligned} \tag{2}$$

Obviously, $\mathbf{G} = V_{\mathrm{tp}} \cup V_{\mathrm{fn}}$; $\mathbf{S} = V_{\mathrm{tp}} \cup V_{\mathrm{fp}}$; $V_{\mathrm{tp}} = \mathbf{G} \cap \mathbf{S}$; and $V_{\mathrm{tp}} \cup V_{\mathrm{fp}} \cup V_{\mathrm{fn}} = \mathbf{G} \cup \mathbf{S}$ where $\cup$ and $\cap$ denote the set union and intersection, respectively. Therefore, it holds that $|\mathbf{G}| = c_{\mathrm{tp}} + c_{\mathrm{fn}}$; $|\mathbf{S}| = c_{\mathrm{tp}} + c_{\mathrm{fp}}$, and $|\mathbf{G} \cup \mathbf{S}| = c_{\mathrm{tp}} + c_{\mathrm{fp}} + c_{\mathrm{fn}}$. The DC [45] and the PVD are defined as

$$\begin{aligned} \mathrm{DC} &= 100 \frac{2c_{\mathrm{tp}}}{2c_{\mathrm{tp}} + c_{\mathrm{fp}} + c_{\mathrm{fn}}} \equiv 100 \frac{2\,|\mathbf{G} \cap \mathbf{S}|}{|\mathbf{G}| + |\mathbf{S}|} \\ \mathrm{PVD} &= 100 \frac{\left(c_{\mathrm{tp}} + c_{\mathrm{fn}}\right) - \left(c_{\mathrm{tp}} + c_{\mathrm{fp}}\right)}{c_{\mathrm{tp}} + c_{\mathrm{fn}}} \equiv 100 \frac{|\mathbf{G}| - |\mathbf{S}|}{|\mathbf{G}|}. \end{aligned} \tag{3}$$

In addition to the DC and PVD, the 95th-percentile bidirectional Hausdorff distance (BHD_{95}) is used to measure dissimilarities between the $\mathbf{G}$ and $\mathbf{S}$ boundaries; see Figure 5. The HD from $\mathbf{G}$ to $\mathbf{S}$ is the maximum distance from the points g from $\mathbf{G}$ to their closest points s in $\mathbf{S}$ [46]:

$$\mathrm{HD}_{G \to S} = \max_{g \in \mathbf{G}} \left\{\min_{s \in \mathbf{S}} d\left(g, s\right)\right\}, \tag{4}$$

where $d(g, s)$ is the Cartesian distance between two 3D points. The HD is asymmetric, as generally $\mathrm{HD}_{G \to S} \neq \mathrm{HD}_{S \to G}$. The symmetric BHD between these two sets is defined as

$$\mathrm{HD}_{\mathbf{G} \leftrightarrow \mathbf{S}} = \max\left\{\mathrm{HD}_{\mathbf{G} \to \mathbf{S}}, \mathrm{HD}_{\mathbf{S} \to \mathbf{G}}\right\}. \tag{5}$$

To decrease the sensitivity to outliers, the 95th-percentile BHD is used in this paper to measure the segmentation accuracy.

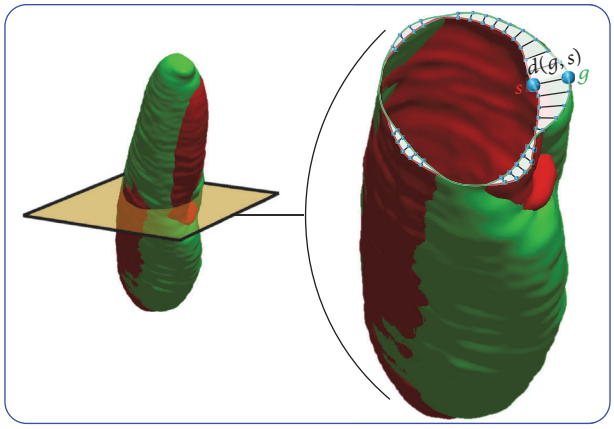


Figure 5: Schematic illustration for the calculation of the Haussdorf distance between the ground truth (green) and segmented (red) objects.

4. Experimental Results

Performance assessment of our framework is carried using dynamic CT kidney data, which were collected from 20 subjects. Each subject dataset consists of three 3D CT scans obtained at the pre- and postcontrast medium administration, namely, noncontrast, postcontrast, and late contrast 3D scan. The CT data were obtained using a GE light speed plus scanner (General Electric, Milwuakee, USA). The CT data acquisition parameters were 120 KV, 250 mA, in-plane resolution: $0.64 \times 0.64\,\text{mm}^2$, slice thickness: 0.9 mm, field-of-view (FOV): 360 mm, the 3D image sizes range from $512 \times 512 \times 232$ to $512 \times 512 \times 366$. In order to minimize the effect of interobserver variability, two experts delineated the kidney borders independently on the CT images and the ground truth labels were considered as the common segmented region of their delineations.

Quantitative evaluation is performed using a leave-one-subject-out approach and the number of decision trees was set to 400. First, all the 3D CT scans (60 scans in total) from all of the 20 subjects are coregistered using our registration methodology described in Section 2.2. To segment a test subject, all of its pre- and postcontrast scans are removed from the training database. Then, the 19 NCC-top-ranked scans are selected from the remaining training scans to build the test scan adaptive shape prior, described by (1) and the method in [39, 40]. Lastly, all regional features described in Sections 2.1 and 2.3 are extracted for (i) the NCC-selected scans to build the training model of the RF; and (ii) the 3D coregistered test scan to be classified using the built RF model. The above steps are repeated for all of the 60 CT volumes of the 20 subjects.

Cross-sectional segmentation results in the axial, sagittal, and coronal views using our technique are demonstrated in Figure 6 for CT data from four subjects at different contrast-enhancement phases. The 3D kidney surface is constructed by accounting for the object labels in the output of the RF classifier. Followed by a postprocessing step using a 3D median filter to smooth the noisy output labels of the classifier. The segmentation accuracy of our framework is assessed using the evaluation metrics described in Section 3. The overall accuracy for all subjects in terms of mean and standard deviation is summarized in Table 1.

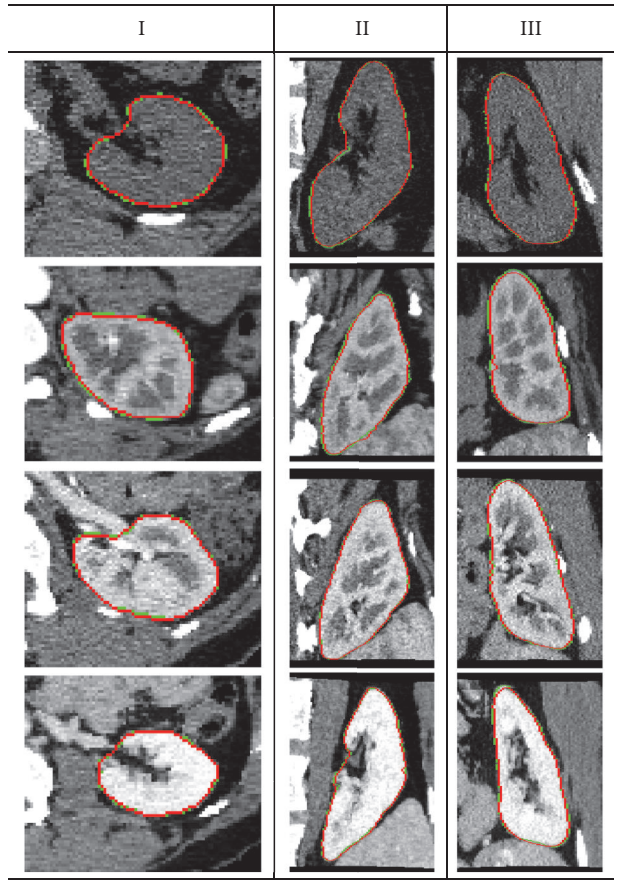


Figure 6: Cross-sectional axial (I), sagittal (II), and coronal (III) segmentation results of our approach for multiple subjects at different contrast-enhancement phases, showing reliable determination of kidney borders (red) compared with the ground truth (green) contours.

Table 1: Segmentation accuracy of our method compared with Zhang et al. [47] approach based on the DC, PVD, and BHD_{95} metrics. Note that DC, PVD, BHD_{95}, and SD stand for Dice coefficient, percentage volume difference, bidirectional 95th-percentile Hausdorff distance, and standard deviation, respectively.

Metric	Segmentation method		p value
	Our	Zhang et al. [47]	
	Mean ± SD	Mean ± SD	
DC (%)	97.27 ± 0.83	91.60 ± 2.29	$\leq 10^{-4}$
BHD_{95} (mm)	0.93 ± 0.49	5.36 ± 1.12	$\leq 10^{-4}$
PVD (%)	2.92 ± 2.21	5.00 ± 3.28	$\leq 10^{-4}$

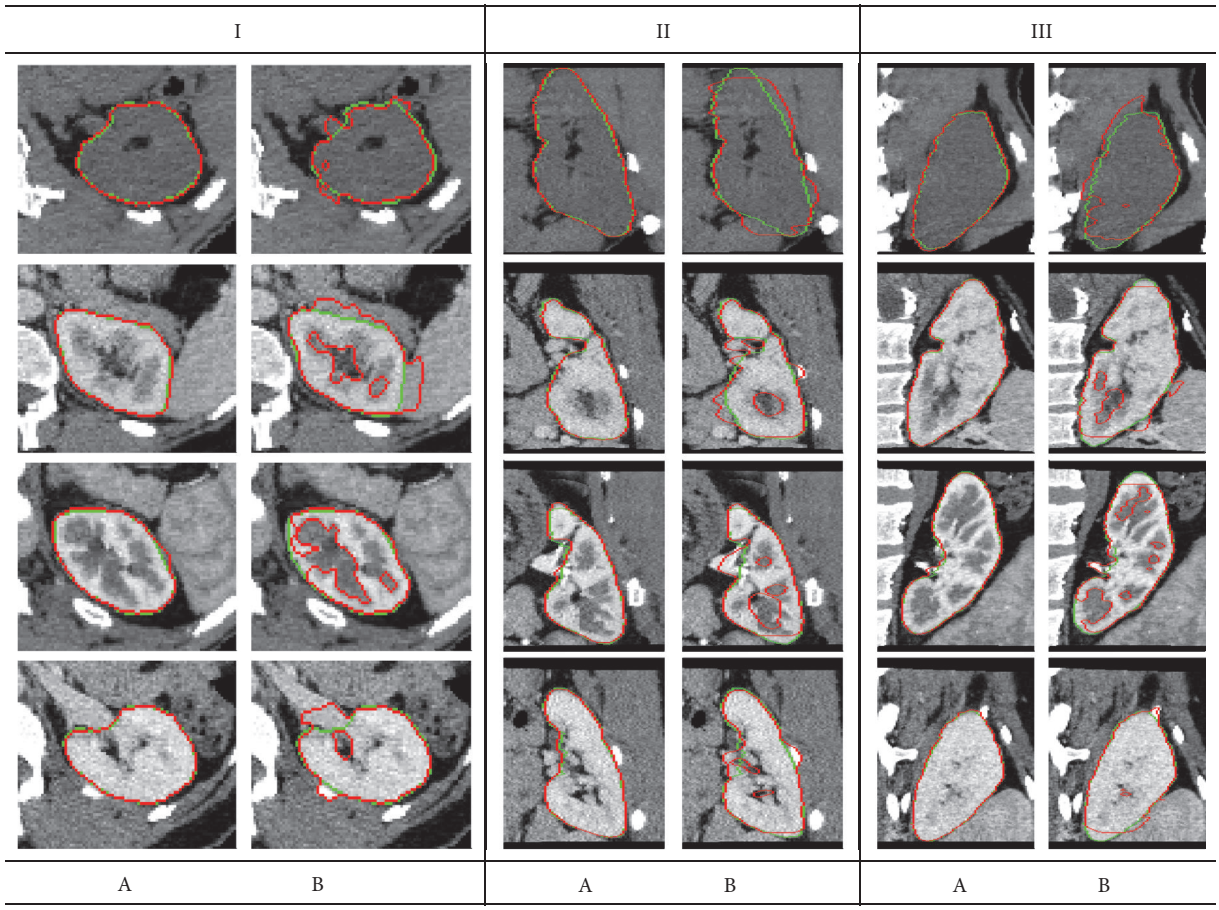


Figure 7: Cross-sectional axial (I), sagittal (II), and coronal (III) segmentation results from multiple subjects at different contrast-enhancement phases of our approach (A) and the approach proposed by Zhang et al. [47] (B). The red and green contours refer to model segmentation and the ground truth, respectively.

In order to demonstrate the high accuracy of our kidney segmentation framework, we compare it with the image segmentation method that was proposed by Zhang et al. [47], which has a freely available software package and thus avoids reimplementing an existing method. Figure 7 demonstrates sample segmentation results comparing our method versus the approach proposed in [47] on multiple subjects. The results in Figure 7 show reliable determination of the kidney borders of our technique compared to Zhang et al. [47] method. Additionally, a summary of the overall segmentation accuracy of our and Zhang et al. [47] methods, with respect to the ground truth delineation, for all data sets, is given in Table 1. According to the higher DC and lower HD_{95} and PVD values in Table 1, our technique performs notably better compared with [47]. This has been documented using the statistical significance of the statistical paired t-test as shown in Table 1 (p value is < 0.05).

In addition to the segmentation evaluation metrics described in Section 2.2, the robustness of our segmentation framework is assessed using the receiver operating characteristics (ROC) [48] as an alternate metric to evaluate the performance of segmentation systems. Generally, the ROC analysis assesses the sensitivity of a segmentation method relative to the choice of its operating point (e.g., a classification threshold). This is achieved by plotting the relationship between the true positive and false positive rates for different operating points. Figure 8 shows the ROC curves of our method and Zhang et al. [47] approach. The figure clearly demonstrates that our technique attained higher performance compared with [47], as evidenced by the area under the ROC curve (AUC) of 0.96 compared with 0.92 for Zhang et al. approach [47].

5. Conclusions

In conclusion, a random forests-based framework is proposed for 3D kidney segmentation from dynamic contrast enhanced abdominal CT images. In order to account for large kidney inhomogeneity and nonlinear intensity variation of the dynamic CT data, our framework integrated two spatial-appearance features, namely, the higher-order spatial interactions features and appearance-based adaptive shape prior features, in addition to the Hounsfield appearance features. Qualitative and quantitative evaluation results confirmed reliable kidney tissue segmentation using our approach at different contrast-enhancement phases of agent transit. This

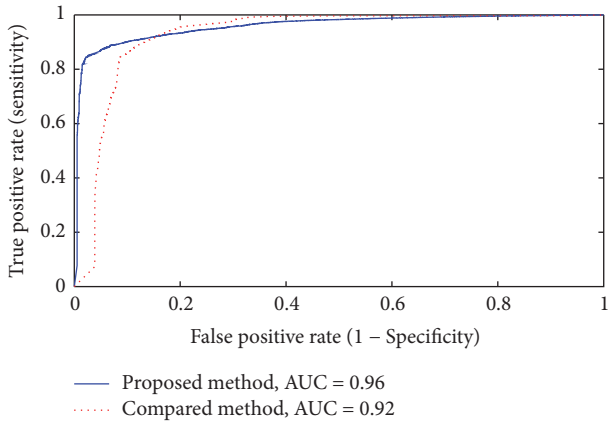


Figure 8: The ROC curves for our segmentation method and the method proposed in [47]. The "AUC" stand for the area under the curve.

has been evaluated on CT data sets collected from 20 subjects using both volumetric and distance-based evaluation metrics. In the future work we will investigate the addition of other features (e.g., scale space, local binary patterns). Also, we plan to test our framework on larger data sets to assess its accuracy, robustness, and limitation. Ultimately, we plan to include this segmentation approach into a kidney-dedicated CAD system for early detection of acute renal transplant rejection and treatment planning.

Competing Interests

The authors declare that they have no competing interests.

References

[1] X. Chen, R. M. Summers, and J. Yao, "Automatic 3D kidney segmentation based on shape constrained GC-OAAM," in *Proceedings of the Medical Imaging 2011: Image Processing*, Lake Buena Vista, Fla, USA, February 2011.

[2] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation," *Annual Review of Biomedical Engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[3] F. Khalifa, A. El-Baz, G. Gimel'farb, R. Ouseph, and M. A. El-Ghar, "Shape-appearance guided level-set deformable model for image segmentation," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR '10)*, pp. 4581–4584, Istanbul, Turkey, August 2010.

[4] E. Widjaja, J. W. Oxtoby, T. L. Hale, P. W. Jones, P. N. Harden, and I. W. McCall, "Ultrasound measured renal length versus low dose CT volume in predicting single kidney glomerular filtration rate," *The British Journal of Radiology*, vol. 77, no. 921, pp. 759–764, 2004.

[5] S.-W. Yoo, J.-S. Cho, S.-M. Noh, K.-S. Shin, and J.-W. Park, "Organ segmentation by comparing of gray value portion on abdominal CT image," in *Proceedings of the 5th International Conference on Signal Processing*, pp. 1201–1208, Beijing, China, August, 2000.

[6] S.-J. Kim, S.-W. Yoo, S.-H. Kim, J.-C. Kim, and J.-W. Park, "Segmentation of kidney without using contrast medium on abdominal CT image," in *Proceedings of the 5th International Conference on Signal Processing*, vol. 2, pp. 1147–1152, Beijing, China, August 2000.

[7] R. Pohle and K. Toennies, "A new approach for model-based adaptive region growing in medical image analysis," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, pp. 238–246, Warsaw, Poland, September 2001.

[8] R. Pohle and K. D. Toennies, "Self-learning model-based segmentation of medical images," *Image Processing & Communications*, vol. 7, no. 3-4, pp. 97–113, 2001.

[9] V. Leonardi, V. Vidal, J.-L. Mari, and M. Daniel, "3D reconstruction from CT-scan volume dataset application to kidney modeling," in *Proceedings of the 27th Spring Conference on Computer Graphics (SCCG '11)*, pp. 111–120, ACM, Viničné, Slovakia, April 2011.

[10] B. Tsagaan, A. Shimizu, H. Kobatake, and K. Miyakawa, "An automated segmentation method of kidney using statistical information," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002: 5th International Conference Tokyo, Japan, September 25–28, 2002 Proceedings, Part I*, vol. 2488 of *Lecture Notes in Computer Science*, pp. 556–563, Springer, Berlin, Germany, 2002.

[11] X. Wang, L. He, and W. Wee, "Deformable contour method: a constrained optimization approach," *International Journal of Computer Vision*, vol. 59, no. 1, pp. 87–108, 2004.

[12] J. Lu, J. Chen, J. Zhang, and W. Yang, "Segmentation of kidney using CV model and anatomy priors," in *Proceedings of the Medical Imaging, Parallel Processing of Images, and Optimization Techniques (MIPPR '07)*, Proceedings of SPIE, pp. 678–911, Wuhan, China, November 2007.

[13] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[14] Y.-P. Huang, P.-C. Chung, C.-L. Huang, and C.-R. Huang, "Multiphase level set with multi dynamic shape models on kidney segmentation of CT image," in *Proceedings of the IEEE Biomedical Circuits and Systems Conference (BioCAS '09)*, pp. 141–144, Beijing, China, November 2009.

[15] P. Campadelli, E. Casiraghi, and S. Pratissoli, "A segmentation framework for abdominal organs from CT scans," *Artificial Intelligence in Medicine*, vol. 50, no. 1, pp. 3–11, 2010.

[16] F. Khalifa, A. Elnakib, G. M. Beache et al., "3D kidney segmentation from CT images using a level set approach guided by a novel stochastic speed function," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2011*, pp. 587–594, Springer, 2011.

[17] F. Khalifa, G. Gimel'farb, M. Abo El-Ghar et al., "A new deformable model-based segmentation approach for accurate extraction of the kidney from abdominal CT images," in *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP '11)*, pp. 3393–3396, September 2011.

[18] M. Freiman, A. Kronman, S. J. Esses, L. Joskowicz, and J. Sosna, "Nonparametric iterative model constraint graph min-cut for automatic kidney segmentation," in *Proceedings of the 13th International Conference on Medical Image Computing and Computer-assisted Intervention: Part III (MICCAI '10)*, pp. 73–80, Beijing, China, September 2010.

[19] D.-T. Lin, C.-C. Lei, and S.-W. Hung, "Computer-aided kidney segmentation on abdominal CT images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 59–65, 2006.

[20] M. Spiegel, D. A. Hahn, V. Daum, J. Wasza, and J. Hornegger, "Segmentation of kidneys using a new active shape model generation technique based on non-rigid image registration," *Computerized Medical Imaging and Graphics*, vol. 33, no. 1, pp. 29–39, 2009.

[21] X. Chen, J. K. Udupa, U. Bagci, Y. Zhuge, and J. Yao, "Medical image segmentation by combining graph cuts and oriented active appearance models," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2035–2046, 2012.

[22] R. Cuingnet, R. Prevost, D. Lesage, L. D. Cohen, B. Mory, and R. Ardon, "Automatic detection and segmentation of kidneys in 3D CT images using random forests," *Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 15, no. 3, pp. 66–74, 2012.

[23] H. Badakhshannoory and P. Saeedi, "A model-based validation scheme for organ segmentation in CT scan volumes," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 9, pp. 2681–2693, 2011.

[24] U. Bagci, X. Chen, and J. K. Udupa, "Hierarchical scale-based multiobject recognition of 3-D anatomical structures," *IEEE Transactions on Medical Imaging*, vol. 31, no. 3, pp. 777–789, 2012.

[25] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1723–1730, 2013.

[26] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation from CT images using conditional shape-location and unsupervised intensity priors," *Medical Image Analysis*, vol. 26, no. 1, pp. 1–18, 2015.

[27] G.-Y. Dai, Z.-C. Li, J. Gu, L. Wang, X.-M. Li, and Y.-Q. Xie, "Segmentation of kidneys from computed tomography using 3D fast growcut algorithm," *Applied Mechanics and Materials*, vol. 333–335, pp. 1145–1150, 2013.

[28] E. Zhao, Y. Liang, and H. Fan, "Contextual information-aided kidney segmentation in CT sequences," *Optics Communications*, vol. 290, pp. 55–62, 2013.

[29] C. Chu, M. Oda, T. Kitasaka et al., "Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*, pp. 165–172, Springer, 2013.

[30] G. Yang, J. Gu, Y. Chen et al., "Automatic kidney segmentation in CT images based on multi-atlas image registration," in *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '14)*, pp. 5538–5541, August 2014.

[31] J. Liu, M. G. Linguraru, S. Wang, and R. M. Summers, "Automatic segmentation of kidneys from non-contrast CT images using efficient belief propagation," in *SPIE Proceedings of the Medical Imaging: Computer-Aided Diagnosis*, vol. 8670, Orlando, Fla, USA, February 2013.

[32] J. Liu, S. Wang, M. G. Linguraru, J. Yao, and R. M. Summers, "Computer-aided detection of exophytic renal lesions on non-contrast CT images," *Medical Image Analysis*, vol. 19, no. 1, pp. 15–29, 2015.

[33] F. Khalifa, A. Soliman, A. C. Dwyer, G. Gimel'farb, and A. El-Baz, "A random forest-based framework for 3D kidney segmentation from dynamic contrast-enhanced CT images," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '16)*, pp. 3399–3403, IEEE, Phoenix, Ariz, USA, September 2016.

[34] A. A. Farag, A. S. El-Baz, and G. Gimel'farb, "Precise segmentation of multimodal images," *IEEE Transactions on Image Processing*, vol. 15, no. 4, pp. 952–968, 2006.

[35] 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.

[36] T. K. Ho, "Random decision forests," in *Proceedings of the International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282, Montreal, Canada, August 1995.

[37] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[38] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How mantrees in a random forest?" in *Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 154–168, 2012.

[39] A. Soliman, F. Khalifa, A. Elnakib et al., "Accurate lungs segmentation on CT chest images by adaptive appearance-guided shape modeling," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 263–276, 2017.

[40] F. Khalifa, A. Soliman, A. Takieldeen et al., "Kidney segmentation from CT images using a 3D NMF-guided active contour model," in *Proceedings of the IEEE 13th International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '16)*, pp. 432–435, Prague, Czech Republic, April 2016.

[41] F. Khalifa, G. M. Beache, G. Gimelfarb, J. S. Suri, and A. El-Baz, "State-of-the-art medical image registration methodologies: a survey," in *Handbook of Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, vol. 1, chapter 9, pp. 235–280, Springer, New York, NY, USA, 2011.

[42] B. Glocker, A. Sotiras, N. Komodakis, and N. Paragios, "Deformable medical image registration: setting the state of the art with discrete methods," *Annual Review of Biomedical Engineering*, vol. 13, pp. 219–244, 2011.

[43] F. Y. Wu, "The Potts model," *Reviews of Modern Physics*, vol. 54, no. 1, pp. 235–268, 1982.

[44] F. Khalifa, G. M. Beache, M. A. El-Ghar et al., "Dynamic contrast-enhanced MRI-based early detection of acute renal transplant rejection," *IEEE Transactions on Medical Imaging*, vol. 32, no. 10, pp. 1910–1927, 2013.

[45] K. H. Zou, S. K. Warfield, A. Bharatha et al., "Statistical validation of image segmentation quality based on a spatial overlap index," *Academic Radiology*, vol. 11, no. 2, pp. 178–189, 2004.

[46] G. Gerig, M. Jomier, and M. Chakos, "Valmet: a new validation tool for assessing and improving 3D object segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings*, vol. 2208 of *Lecture Notes in Computer Science*, pp. 516–523, Springer, Berlin, Germany, 2001.

[47] Y. Zhang, B. J. Matuszewski, L.-K. Shark, and C. J. Moore, "Medical image segmentation using new hybrid level-set method," in *Proceedings of the 5th International Conference BioMedical Visualization: Information Visualization in Medical and Biomedical Informatics (MediVis '08)*, pp. 71–76, London, UK, July 2008.

[48] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 5109530, 9 pages
https://doi.org/10.1155/2017/5109530

Research Article

Comparison of Different Features and Classifiers for Driver Fatigue Detection Based on a Single EEG Channel

Jianfeng Hu

Jiangxi University of Technology, Nanchang 330098, China

Correspondence should be addressed to Jianfeng Hu; huguess211@hotmail.com

Received 11 November 2016; Revised 27 December 2016; Accepted 15 January 2017; Published 31 January 2017

Academic Editor: Ayman El-Baz

Driver fatigue has become an important factor to traffic accidents worldwide, and effective detection of driver fatigue has major significance for public health. The purpose method employs entropy measures for feature extraction from a single electroencephalogram (EEG) channel. Four types of entropies measures, sample entropy (SE), fuzzy entropy (FE), approximate entropy (AE), and spectral entropy (PE), were deployed for the analysis of original EEG signal and compared by ten state-of-the-art classifiers. Results indicate that optimal performance of single channel is achieved using a combination of channel CP4, feature FE, and classifier Random Forest (RF). The highest accuracy can be up to 96.6%, which has been able to meet the needs of real applications. The best combination of channel + features + classifier is subject-specific. In this work, the accuracy of FE as the feature is far greater than the Acc of other features. The accuracy using classifier RF is the best, while that of classifier SVM with linear kernel is the worst. The impact of channel selection on the Acc is larger. The performance of various channels is very different.

1. Introduction

Traffic accidents are more and more increasing, resulting in a very large number of casualties. Safety driving is fundamental to public health, and fatigue driving can be life threatening. It is crucial and necessary to develop some technologies for detecting driver fatigue [1–3]. There are many methods that have been proposed in the past few years, such as vehicle driving parameters by using various sensors [4], driver behavior characteristics by using video imaging techniques [5, 6], driver physiological parameters by using acquisition and analysis of electrocardiogram (ECG) [7], electrooculogram (EOG) [8], electromyogram (EMG) [9], and EEG [10–12]. As a kind of direct indicator of the brain status, EEG is considered as the "gold" method to identify driver fatigue.

EEG is an objective method for the evaluation of brain state and function, which is often used in auxiliary diagnosis of illness such as epilepsy and seizure. The advantages of EEG are sensitivity for analysis and being relatively cheap for acquisition. Various computational approaches based on EEG signals have been developed for analyzing and detecting driver fatigue.

Fu et al. [13] proposed a fatigue detection model based on Hidden Markov Model and fused physiological and contextual knowledge to assess probabilities of fatigue. They achieved highest accuracy of 92.5% based on EEG signals from two channels (O1 and O2) and other physiological signals. Li et al. [14] collected 16 channels of EEG data and computed 12 types of energy parameters. The number of significant electrodes is reduced using Kernel Principle Component Analysis (KPCA). The experimental results from two channels (FP1 and O1) achieved the highest accuracy of 91.5%. Wali et al. [15] used Discrete Wavelet Transforms to process the EEG signal for fatigue detection and yielded the highest accuracy of 85%. Using Fast Fourier Transform, Simon et al. [16] proposed EEG alpha spindle measures for assessing driver fatigue. Charbonnier et al. [17] made use of the Frobenius distance between the EEG spatial covariance matrices of 6 brain regions, and experimental results had shown that the index based on the alpha band can accurately assess fatigue. Apker et al. [18] predicted driver performance using power spectral density and the linear regression, providing a confidence estimate for the stable driving model. Hajinoroozi et al.'s experimental results showed that channel-wise

convolutional neural network achieved robust and improved performance for detection of driver fatigue [19]. Zhao et al. [20] studied an automatic measurement of driving mental fatigue, using a KPCA-SVM classifier and their accuracy was quite high, up to 98.7%. Kong et al. [21] analyzed EEG signals by using Granger-Causality-based brain effective networks and found a significant difference in terms of strength of Granger-Causality in the frequency domain and some changes were more significant over the frontal brain. Zhao et al. [22] observed that coherence was significantly increased in the frontal, central, and temporal brain regions, as well as significant increases in the clustering coefficient and the character path length.

Recently, entropy has been broadly applied in the analysis of EEG signals, considering the fact that it is a complex, unstable, and nonlinear signal [23–28]. Xiong et al. combined features of AE and SE with Support Vector Machine (SVM) classifier to detect driver fatigue, achieving highest accuracy of 91.3% at channel P3 [25]. Chai et al. present independent component by entropy rate bound minimization analysis for the source separate, autoregressive (AR) modeling for the features extraction and Bayesian neural network for the classification algorithm. They achieved an accuracy of 88.2% and the highest value of area under the receiver operating curve (AUC) is 0.93 [26]. Zhang et al. extracted wavelet entropy and SE of EEG and wavelet entropy of EOG and AE of EMG to estimate the driving fatigue stages, and their accuracy was quite high, which is about 96.5%–99.5% using artificial neural network [27]. Kar et al. used five types of entropies, that is, Shannon's entropy, Rényi entropy of order 2, Rényi entropy of order 3, Tsallis wavelet entropy, and Generalized Escort-Tsallis entropy, along with alpha band relative energy for estimation of fatigue level [28]. However, few studies have been conducted for using optimal combination of entropy methods and classifiers based on EEG to study driver fatigue detection.

Multichannels EEG acquisition system, such as the 32-channel EEG system used in my experiment, is relatively complex equipment, which can only be available in laboratories or hospitals. It requires well-trained technicians to locate electrodes, since all the electrodes have to be placed in the proper location. And it is time-consuming. All these reasons are making the system difficult to apply in real life. Therefore, a worthwhile EEG system with fewer channels or even one channel for estimating driver fatigue has to be a portable system that is cheaper, simpler, and easier to use.

Although many EEG-based methods have been proven to detect driver fatigue, the optimal method has not yet been determined. Furthermore, the EEG with more channels usually restricts its application in the detection of driver fatigue. Using the data from 12 subjects, the detection model for driver fatigue was developed with a single channel. Four types of entropies were deployed in this work: SE, FE, AE, and PE. The classification procedure was implemented by ten classifiers: K-Nearest Neighbors (KNN), SVM with linear kernel (LS), SVM with RBF kernel (RS), Gaussian Process (GP), Decision Tree (DT), RF, Multilayer Perceptron (MLP), AdaBoost (AB), Gaussian Naïve Bayes (GNB), and Quadratic Discriminant Analysis (QDA). The aims of the present study are to determine the optimal combination of feature, classifier, and channel that can be effective in portable application with a single channel.

Figure 1: Snapshot of the experimental setup.

The rest of the paper is organized as follows: Section 2 describes the proposed methodology. Results and discussion are reported in Section 3. Conclusion is reported in Section 4.

2. Materials and Methods

2.1. Subjects. Twelve university students (men, 19–24 years) participated in this experiment. All the subjects were asked to be out of any type of stimulus like alcohol, medicine, tea, or coffee before and during the experiment. Before the experiment, they practiced the driving task for several minutes to become acquainted with the experimental procedures and purposes. All experimental procedures were performed using a static driving simulator in a software-controlled environment. This work was approved by Academic Ethics Committee of Jiangxi University of Technology.

2.2. Experiment. The experimental setup of the work is based on our previous work. A sustained-attention driving task was performed by each subject on a static driving simulator (The ZY-31D car driving simulator, produced by Peking ZhongYu Co., Ltd.) with a wide screen composed of three 24-inch monitors shown as in Figure 1. On the screen, a customized version of the Peking ZIGUANGJIYE software ZG-601 (Car Driving Simulation Teaching System V9.2) was shown. The driving environment selected for this study was a highway with low traffic density and the driving task was started at 9 a.m. After the 5-minute practice session, each subject was given a break of 10 min away from the simulator and was allowed to have unconstrained movement within the laboratory. Then they commenced their about 1-2 hours of driving after a quick check of all instrumentation.

2.3. Data Recording and Preprocessing. First, when the subjects had been driving for 20 minutes, the last 5-minute recorded EEG signal was labeled as normal state; second, when the continuous driving procedure lasted 60–120 minutes until the questionnaire results show the subject was in driving fatigue, obeying Lee's subjective fatigue scale and Borg's CR-10 scale [29, 30], the last 5-minute recorded EEG signal was labeled as fatigue state. All channel data were referenced to two electrically linked mastoids at A1 and A2, digitized at 1000 Hz from a 32-channel electrode cap (including

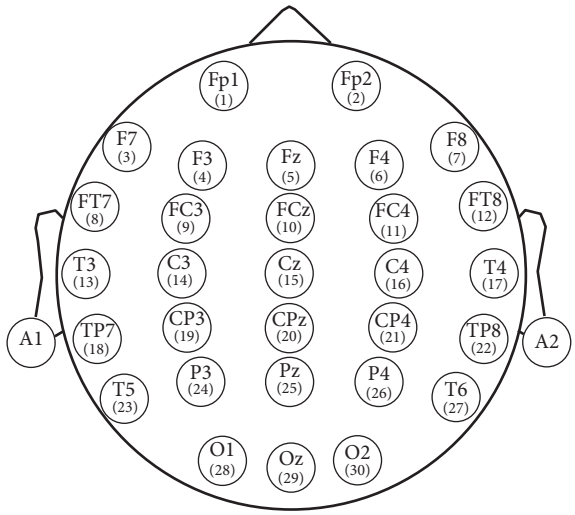


Figure 2: Electrodes position according to International 10-20 System standards.

30 effective channels and 2 reference channels) based on the International 10-20 System (Figure 2).

After the EEG signals acquisition, the main steps of data preprocessing were carried out by the Scan 4.3 software of Neuroscan. The raw signals were first filtered by a 50 Hz notch filter and a 0.15 Hz to 45 Hz band-pass filter was used. Then 5-minute EEG signals from 30 channels were sectioned into 1-s epochs, resulting in 300 epochs. With the 12 subjects, a total of 3600 epochs of dataset were formed for the normal state and another 3600 epochs for the fatigue state.

2.4. Feature Extraction. In recent years, various entropies have been expanded in several different fields [31]. As the nonlinear parameters can quantify the complexity of a time series, it can be used to evaluate the nonlinear, unstable EEG signals [32]. PE is calculated by applying the Shannon function to the normalized power spectrum, and the calculation algorithm is as described in literature [33]. AE, proposed by Pincus [34], is calculated in time domain without phase space reconstruction of the signal. Similar to AE, SE is proposed by Richman and Moorman [35]. The calculation algorithm of AE and SE is defined clearly as described in literature [36]. FE can get stable results for different parameters and offers better noise resistance, defined clearly as described in literature [37].

In the above four types of entropies, AE, SE, and FE have parameters, m and r, which are the dimensions of phase space and similarity tolerance, respectively. Generally, too larger of r will lead to a loss of useful information. However, if r is underestimated, the sensitivity to noise will be increased significantly. In the present study, $m = 2$ while $r = 0.2 * \mathrm{SD}$, where SD denotes the standard deviation of the time series according to literature [38].

For optimizing the detection quality, the features were normalized for each subject by scaling between −1 and 1.

2.5. Classification. Since there is no uniform classification method suitable for all subjects and all applications, usually it may be useful to test multiple methods. In this work, I have used ten classifiers, namely, KNN, LS, RS, GP, DT, RF, MLP, AB, GNB, and QDA. They are briefly explained below.

2.5.1. KNN. Neighbors-based classification does not construct a general model but simply compares instances of features of the training data. KNN is a supervised learning technique where a new instance is classified based on the closest training samples present in the feature space [39]. KNN implements learning based on the k-Nearest Neighbors of each query point, where k is 5 in this study.

2.5.2. SVM. In the case of nonlinear classification, kernels, such as radial basis functions (RBF), are used to map the data into a higher dimensional feature space in which a linear separating hyperplane could be found [40]. When the number of samples is less than the number of features, nonlinear learning methods do not significantly affect the results and it may be better to simply use linear learning method. So SVM with linear kernel (LS) and SVM with RBF kernel (RS) were both chosen as the classifier in this work.

When training an SVM classifier with the RBF kernel, two parameters must be considered: c and γ. A lower c makes the decision surface smooth, while a higher c aims at classifying all training examples correctly. γ defines how much influence a single training example has. In this study, $\gamma = 2$ and $c = 1$.

2.5.3. GP. The GP Classifier implements Gaussian Processes for classification purposes, more specifically for probabilistic classification [41].

2.5.4. DT. DT is a nonparametric supervised learning method used for classification [42]. DT creates a series of binary decisions on the features which best distinguishes classes. The maximum depth of the tree is 10 in this work.

2.5.5. RF. RF fits a number of Decision Tree classifiers on various subdatasets and averages predicted accuracy [43]. In this work, the maximum depth of the tree is 10 and the number of trees in the forest is 10.

2.5.6. MLP. MLP trains using gradient descent and the gradients are calculated using Backpropagation (BP) [44].

2.5.7. AB. AB classifier begins by fitting a classifier on the raw dataset and then fits additional copies of the classifier on the same dataset where the weights of incorrectly classified instances are adjusted [45].

2.5.8. GNB. Naive Bayes method is based on applying Bayes' theorem with the "naive" assumption [46]. The likelihood in GNB of the features is assumed to be Gaussian.

2.5.9. QDA. QDA searches for a linear combination of features which statistically best distinguishes objects in different classes from each other [47]. QDA classifier has a quadratic decision boundary.

2.6. Performance Metrics. For developing a new detector and estimating its potential application performance, it is very

Table 1: Optimal combination for different subjects.

Subject	Optimal combination	Highest Acc	AUC
1	FE + KNN	94.3%	0.976
2	FE + KNN	86.4%	0.929
3	FE + RF	93.4%	0.981
4	FE + RF	91.0%	0.969
5	FE + RF	92.6%	0.976
6	FE + RF	91.3%	0.974
7	FE + RF	91.4%	0.968
8	FE + RF	92.7%	0.981
9	FE + RF	94.4%	0.983
10	FE + RF	91.9%	0.975
11	FE + RF	90.5%	0.967
12	FE + RF	93.2%	0.979

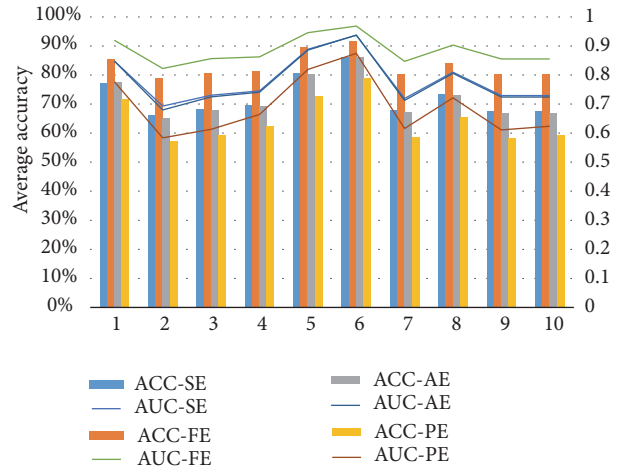


Figure 3: Comparison of performances of four features and ten classifiers. The left s vertical coordinate is for average accuracy (%) for 12 subjects, while the right vertical coordinate is for average AUC for 12 subjects. The horizontal coordinate is for classifier. 1–10 represent KNN, LS, RS, GP, DT, RF, MLP, AB, GNB, and QDA, respectively. ACC-SE, ACC-FE, ACC-AE, and ACC-PE represent accuracy with features SE, FE, AE, and PE, respectively. AUC-SE, AUC-FE, AUC-AE, and AUC-PE represent AUC with features SE, FE, AE, and PE, respectively.

important to examine properly the detection quality [48]. The leave-one-out (LOO) cross-validation approach is used to assess the performance of the system for driver fatigue detection. The total average accuracy based on some feature and the classifier is the average of the accuracy of all single channels based on the same feature and same classifier.

To provide an easier-to-understand method to measure the detection quality, the well-known performance indicators [43], including accuracy (Acc), sensitivity (Sn), and specificity (Sp), were described as follows:

$$\begin{aligned} \mathrm{Sn} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{Sp} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{Acc} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \end{aligned} \tag{1}$$

where TP (true positive) denotes the number of the data inputs that refer to fatigue state correctly classified as fatigue. FP (false positive) is the number of data inputs that refer to normal state classified as fatigue state. TN (true negative) is number of the data inputs that refer to normal state correctly classified as normal state. FN (false negative) is the data inputs that refer to fatigue state classified as normal state.

AUC illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) versus the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as Sn, and FPR is one minus the Sp.

3. Results and Discussion

3.1. Comparison of Performances of Different Subjects. As shown in Figure 3 and Table 1, the best average accuracy is produced in combination of FE + RF (where average accuracy is 91.7%) and the worst average accuracy is produced in combination of SE + LS (where average accuracy is 57.4%). It can be found that the best accuracies of Subject 1 and Subject 2 all occurred in the combination of FE + KNN while, for Subjects 3–12, best recognition rates all appear in the combination of FE + RF. The worst recognition rate of Subject 1 appears in combination of SE + LS, while, for Subjects 2, 5, 6, 7, 9, 11, and 12, it appears in the combination of PE + LS, and, for Subjects 3, 4, 8, and 10, it appears in the combination of PE + MLP. For all 12 subjects, the highest accuracy is 94.4% for Subject 9 with the combination of FE + RF, and the worst recognition rate (51.7%) also appeared in Subject 9 with the combination of SE + LS. This is an interesting phenomenon. For the same subject, using different methods, some subjects will have a particularly larger difference, and some may be less.

As for the AUC, there are similar results. The best AUC is produced in combination of FE + RF (where average AUC is 0.969) and the worst average accuracy is produced in combination of PE + LS (where average AUC is 0.584). For all 12 subjects, the highest AUC (0.983) appears in Subject 9, and the worst AUC (0.517) also appears in Subject 9. This is very similar to ACC.

Different subjects have different brain characteristics, so the EEG features are different. Different subjects using the same feature extraction method or the same classifier may have different performances. The result has two meanings, one is that it is possible to choose a combination that is subject-specific, which is different from the subjects using different combination, thus improving the recognition rate of each subject. Two is that subject-specific EEG feature can be distinguished from different subjects for identification or authentication of individual, that is, the EEG password or biometrics [49, 50].

Table 2: Comparison of mean accuracy (%) of combination of four features and ten classifiers.

Classifier	Feature				
	SE	FE	AE	PE	Mean ± SD
AB	73.2 ± 4.4	84.2 ± 3.6	72.9 ± 5.5	65.3 ± 6.1	73.9 ± 8.4
DT	80.6 ± 3.3	89.7 ± 2.9	80.2 ± 4.2	72.7 ± 5.7	80.8 ± 7.3
GP	69.5 ± 5.4	81.7 ± 4.1	69.0 ± 6.4	62.8 ± 6.0	70.8 ± 8.8
LS	66.0 ± 5.0	79.3 ± 4.4	64.8 ± 6.4	57.3 ± 7.2	66.9 ± 9.9
GNB	67.5 ± 6.0	80.5 ± 4.3	66.8 ± 7.1	58.2 ± 7.1	68.3 ± 10.1
KNN	77.3 ± 4.4	85.8 ± 3.4	77.4 ± 5.0	71.5 ± 7.6	78.0 ± 7.4
MLP	67.7 ± 5.5	80.7 ± 4.3	67.0 ± 6.8	58.5 ± 7.2	68.4 ± 10.0
QDA	67.5 ± 6.0	80.5 ± 4.3	66.8 ± 7.1	59.3 ± 7.1	68.5 ± 9.8
RF	85.9 ± 3.1	**91.8 ± 2.7**	85.9 ± 3.3	79.1 ± 9.3	85.7 ± 7.0
RS	68.3 ± 5.7	81.2 ± 4.1	67.9 ± 6.8	59.2 ± 6.8	69.1 ± 9.8
Mean ± SD	72.3 ± 8.1	83.5 ± 5.6	71.9 ± 9.0	64.4 ± 10.1	

Boldface indicates FE + RF is the optimal method.

Table 3: Comparison of mean AUC of combination of four features and ten classifiers.

Classifier	Feature				
	SE	FE	AE	PE	Mean ± SD
AB	0.808 ± 0.044	0.904 ± 0.027	0.804 ± 0.053	0.720 ± 0.080	0.809 ± 0.085
DT	0.886 ± 0.033	0.946 ± 0.025	0.883 ± 0.038	0.817 ± 0.060	0.883 ± 0.061
GP	0.743 ± 0.059	0.865 ± 0.037	0.736 ± 0.069	0.667 ± 0.077	0.753 ± 0.095
LS	0.690 ± 0.053	0.825 ± 0.055	0.674 ± 0.068	0.584 ± 0.098	0.693 ± 0.111
GNB	0.726 ± 0.063	0.857 ± 0.036	0.720 ± 0.073	0.609 ± 0.090	0.728 ± 0.111
KNN	0.847 ± 0.044	0.921 ± 0.025	0.847 ± 0.050	0.775 ± 0.099	0.848 ± 0.080
MLP	0.716 ± 0.063	0.850 ± 0.040	0.709 ± 0.075	0.615 ± 0.092	0.722 ± 0.109
QDA	0.726 ± 0.063	0.857 ± 0.036	0.720 ± 0.073	0.622 ± 0.090	0.731 ± 0.108
RF	0.936 ± 0.031	**0.969 ± 0.021**	0.937 ± 0.031	0.874 ± 0.111	0.929 ± 0.070
RS	0.0728 ± 0.062	0.859 ± 0.036	0.721 ± 0.074	0.610 ± 0.087	0.729 ± 0.111
Mean ± SD	0.780 ± 0.095	0.885 ± 0.057	0.775 ± 0.104	0.689 ± 0.132	

Boldface indicates FE + RF is the optimal method.

3.2. Comparison of Four Feature Methods. From the above results, the combination of entropy and classifier improved the classification performance. Because the main purpose of my study is to find the optimal combination of feature and classifier based on a single EEG channel, in order to evaluate the performance influence on different entropy features, four types of entropy feature methods and ten classifiers were compared. Figure 3 shows the mean accuracy of generated features obtained from the four entropy methods based on EEG signals from all single channels of 12 subjects, using ten classifiers. From Figure 3, I can conclude that the classification accuracy of the combination of FE with any one of the classifiers is better than combination of the other feature methods with any one of the classifiers. Hence FE was selected as best feature in this work as it is robust and efficient. The detector of using FE + RF fusion method could present a better performance and robustness.

As shown in Table 2, the average accuracy was compared with 12 subjects based on different feature and classifier. The average accuracy based on FE feature was 83.5%, while the average accuracy based on PE feature was 64.4%. The highest mean Acc appeared at the combination of FE + RF, reaching 91.8%, while the worst mean Acc appeared at the combination of PE + LS, achieving 57.3%. These results are in agreement with the results of Section 3.1.

As shown in Table 3, the average AUC was compared with 12 subjects based on different feature and classifier. The average AUC based on FE feature was 0.885, while the average AUC based on PE feature was 0.689. The highest mean AUC occurred at the combination of FE + RF, reaching 0.969, while the worst mean AUC occurred at the combination of PE + LS, achieving 0.584. These results are also in agreement with the results of Section 3.1.

3.3. Comparison of Ten Classifiers. Overall, sorting from large to small of the average accuracy of ten classifiers based on four features and 12 subjects is RF\DT\KNN\AB\GP\RS\QDA\MLP\GNB\LS. The sort of mean AUC is the same.

For 12 subjects, I used $k = 1$, 3, and 5 for KNN and found that $k = 5$ gave the best performance. It can be seen that KNN achieves the highest accuracy with 94.3% and AUC of 0.976

Table 4: Studies regarding driver fatigue detection using different types of entropy.

Research group	Feature method	EEG channels	Highest accuracy
Li et al. [14]	12 types of energy parameters	FP1 and O1	91.5%
Zhang et al. [27]	Approximate entropy	O1 and O2	96.5%
Khushaba et al. [51]	Fuzzy entropy	Fz, T8, and Oz	92.8%
Zhao et al. [52]	Sample entropy	F3	95.0%
This paper	Fuzzy entropy	CP4	96.6%

with FE feature. These accuracies are better than previous studies.

3.4. Comparison of Channels. For channel comparison, the performance of each channel is determined. In order to compare the performance of each channel, with average of 12 subjects, the four types of combinations were compared, including combination of the best features and the best classifier (FE + RF), combination of the best feature + the worst classifier (FE + LS), combination of the worst features and the best classifier (PE + RF), combination of the worst feature + classifier (PE + LS). It can be seen that the highest Acc of single channel is 96.6% at the combination of CP4 + FE + RF, which can fully meet the requirements of mobile computing. The worst Acc is only 55.2% at the combination of Cz + PE + LS.

It can be seen from Figure 4 that all channels of the four combinations are sorted according to the Acc. The index is not the same order in the four combinations. For example, the best channel is CP4 at the combination of FE + RF, while the best channel is T6 at the combination of FE + LS, and the best channel is O1 at the combination of both PE + RF and PE + LS, indicating channel selection is related to feature extracted methods and classifier methods largely.

The result of AUC is very similar. The highest AUC of single channel is 0.993 at the combination of CP4 + FE + RF, while the worst AUC is only 0.545 at the combination of Cz + PE + LS.

In addition to the variation of different channels shown as in Figure 4, we are concerned about which part of brain regions these select channels locate over. So the selected electrodes in each subject were mapped onto their corresponding locations in the electrode cap. It can be seen that the distribution of top channels is much more scattered.

The above results demonstrated the system using a single channel could achieve very high accuracy in detecting driver fatigue, while reducing the decisive number of electrodes from 30 to 1. It is possible to use single channel for driver fatigue detection. The highest recognition rate in this work can be up to 96.6%, which is not the worst comparing with other research results.

Sort of channel is not related to hemisphere, and there is no significant correlation between brain areas. For each subject, the best channel is not the same.

For different analysis targets, using different features may have different impacts on the classification accuracy. In this paper I selected four entropies for comparison purpose. Figure 1 indicates that, for the same data source, the classification performances of the four entropies and ten classifiers are notably different. In my experiment paradigm, the combination of feature FE and classifier RF has the highest accuracy if single entropy is used as input. As see in Table 4, it is found that the classification performance of the proposed method was better than the other research using fewer channels of EEG signals; it is expected that the combination of feature FE, classifier RF, and channel CP4 can show better performance for fatigue forecast. Although the present study is based on the existing EEG data, the high performance of detection of driving fatigue by using of FE-based classification indicated well application on the real-time detection of driving fatigue. To realize real-time detection of driving fatigue, I only needed to record a single channel EEG signals when in fatigue state and normal state and then trained FE-based classification. Once the trained classification model is being saved, I could achieve real-time detection of driving fatigue and try to avoid traffic accidents through the alarm.

4. Conclusions

In this paper, an approach based on combination of four entropy features and ten classifiers is proposed to detect driver fatigue in an EEG-based system. Results also showed that it is a promising system to detect driver fatigue, achieving high success rates with only one electrode. The following was found: (1) It is possible to use a single channel for driver fatigue detection. The highest recognition rate in this work can be up to 96.6%, which has been able to meet the needs of real applications. (2) The best combination of channel + features + classifier of different subjects is not the same; that is to say, the best combination is subject-specific. (3) The impact of feature on the accuracy and AUC is larger. In this work, the Acc of FE as the feature is far greater than the Acc of PE as the feature. (4) The impact of the classifier on the Acc and AUC is larger. In this work, the Acc of classifier RF is the best, while classifier LS is the worst. (5) The impact of channel selection on the Acc and AUC is significant.

However, some limitations of this study are as follows: (1) the sample size was relatively small. To extend my research, in the future, I will increase the number of subjects to improve the validation of results and to classify more fatigue states such as severe fatigue. (2) The parameters of classifier did not carry out optimization, such as MLP and SVM which are very sensitive to parameters. It is also possible that there are no optimization parameters, so the performance for classifier MLP and SVM is not good. (3) In this work, only four kinds of entropy feature were compared, no more feature extraction methods, such as AR, wavelet, and spectrum.

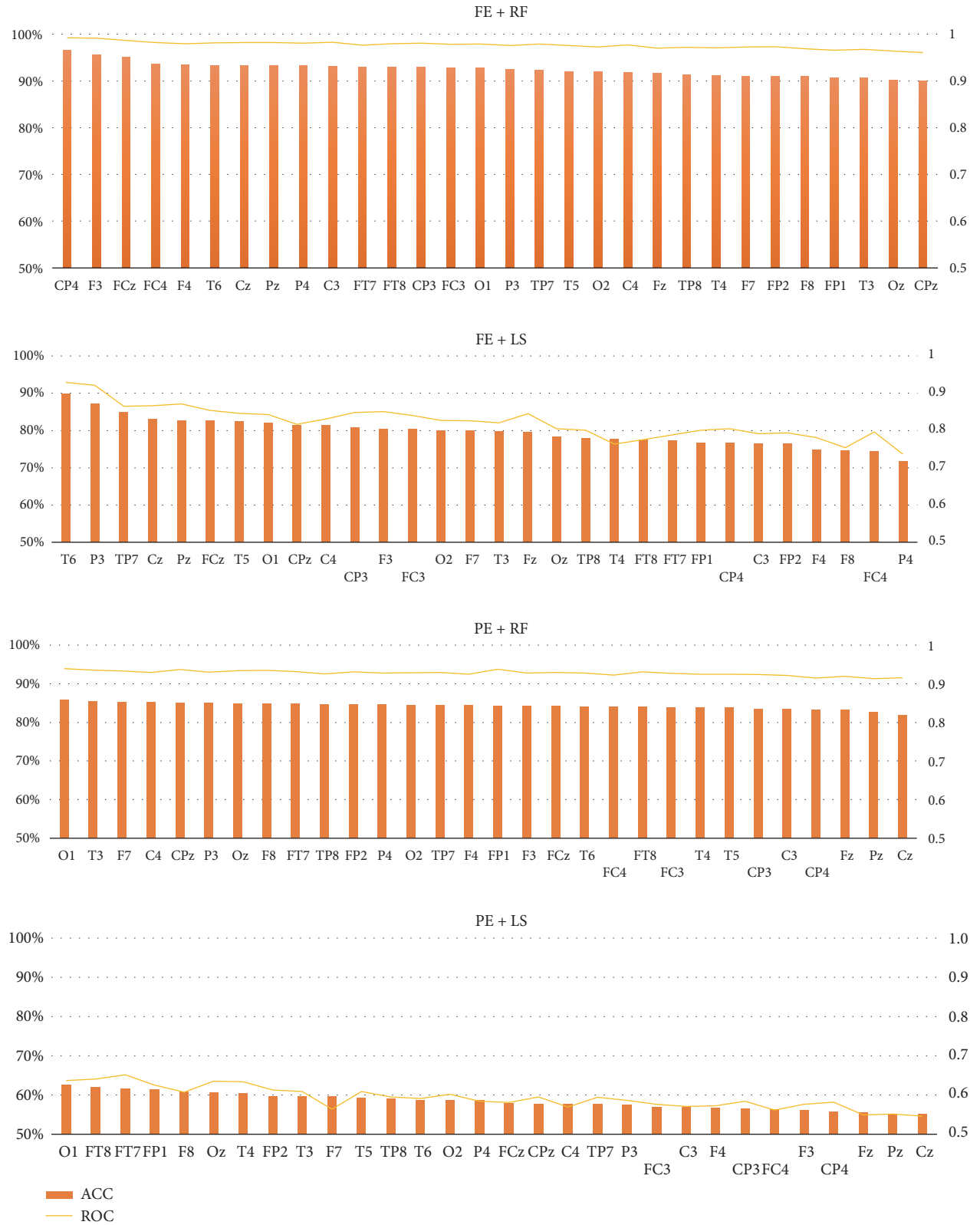


Figure 4: Comparison of all channels based on feature FE and classifier RF of 12 subjects. The left s vertical coordinate is for accuracy (%), while the right vertical coordinate is for AUC. The horizontal coordinate is for channel.

It is hoped that these findings may have the generalizability to provide an effective approach for auxiliary diagnosis of driver fatigue, in order to maintain public health and avoid life threatening.

Competing Interests

The author declares no conflict of interests.

Acknowledgments

This work was supported by Science and Technology Key Project of Jiangxi Provincial Department of Education (GJJ151146) and Patent Transformation Project of Intellectual Property Office of Jiangxi Province (The Application and Popularization of the Digital Method to Distinguish the Direction of Rotation Photoelectric Encoder in Identification). Thanks are due to JL Min and P Wang for collecting and preprocessing EEG data.

References

[1] S. K. L. Lal and A. Craig, "A critical review of the psychophysiology of driver fatigue," *Biological Psychology*, vol. 55, no. 3, pp. 173–194, 2001.

[2] V. Saini and R. Saini, "Driver drowsiness detection system and techniques: a review," *Computer Science and Information Technologies*, vol. 5, no. 3, pp. 4245–4249, 2014.

[3] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: a review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 596–614, 2011.

[4] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: a review," *Sensors*, vol. 12, no. 12, pp. 16937–16953, 2012.

[5] J. Jo, S. J. Lee, K. R. Park, I.-J. Kim, and J. Kim, "Detecting driver drowsiness using feature-level fusion and user-specific classification," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1139–1152, 2014.

[6] G. Niu and C. Wang, "Driver fatigue features extraction," *Mathematical Problems in Engineering*, vol. 2014, Article ID 860517, 10 pages, 2014.

[7] R. Fu and H. Wang, "Detection of driving fatigue by using noncontact emg and ecg signals measurement system," *International Journal of Neural Systems*, vol. 24, no. 3, Article ID 1450006, 15 pages, 2014.

[8] J. Ma, L. Shi, and B. Lu, "An EOG-based vigilance estimation method applied for driver fatigue detection," *Neuroscience & Biomedical Engineering*, vol. 2, no. 1, pp. 41–51, 2014.

[9] H. Wang, "Detection and alleviation of driving fatigue based on EMG and EMS/EEG using wearable sensor," in *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare—"Transforming Healthcare through Innovations in Mobile and Wireless Technologies"*, London, UK, October 2015.

[10] A. G. Correa, L. Orosco, and E. Laciar, "Automatic detection of drowsiness in EEG records based on multimodal analysis," *Medical Engineering & Physics*, vol. 36, no. 2, pp. 244–249, 2014.

[11] Z. Mu, J. Hu, and J. Yin, "Driving fatigue detecting based on EEG signals of forehead area," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2016, Article ID 1750011, 12 pages, 2016.

[12] J. Yin, J. Hu, and Z. Mu, "Developing and evaluating a mobile driver fatigue detection network based on electroencephalograph signals," *Healthcare Technology Letters*, 2016.

[13] R. Fu, H. Wang, and W. Zhao, "Dynamic driver fatigue detection using hidden Markov model in real driving condition," *Expert Systems with Applications*, vol. 63, pp. 397–411, 2016.

[14] W. Li, Q.-C. He, X.-M. Fan, and Z.-M. Fei, "Evaluation of driver fatigue on two channels of EEG data," *Neuroscience Letters*, vol. 506, no. 2, pp. 235–239, 2012.

[15] M. K. Wali, M. Murugappan, and B. Ahmmad, "Wavelet packet transform based driver distraction level classification using EEG," *Mathematical Problems in Engineering*, vol. 2013, Article ID 297587, 10 pages, 2013.

[16] M. Simon, E. A. Schmidt, W. E. Kincses et al., "Eeg alpha spindle measures as indicators of driver fatigue under real traffic conditions," *Clinical Neurophysiology*, vol. 122, no. 6, pp. 1168–1178, 2011.

[17] S. Charbonnier, R. N. Roy, S. Bonnet, and A. Campagne, "EEG index for control operators' mental fatigue monitoring using interactions between brain regions," *Expert Systems with Applications*, vol. 52, pp. 91–98, 2016.

[18] G. Apker, B. Lance, S. Kerick, and K. McDowell, "Combined linear regression and quadratic classification approach for an EEG-based prediction of driver performance," in *Proceedings of the International Conference on Augmented Cognition*, pp. 231–240, Springer, Las Vegas, Nev, USA, July 2013.

[19] M. Hajinoroozi, Z. Mao, T. Jung, C. Lin, and Y. Huang, "EEG-based prediction of driver's cognitive performance by deep convolutional neural network," *Signal Processing: Image Communication*, vol. 47, pp. 549–555, 2016.

[20] C. Zhao, C. Zheng, M. Zhao, J. Liu, and Y. Tu, "Automatic classification of driving mental fatigue with EEG by wavelet packet energy and KPCA-SVM," *International Journal of Innovative Computing, Information and Control*, vol. 7, no. 3, pp. 1157–1168, 2011.

[21] W. Kong, W. Lin, F. Babiloni, S. Hu, and G. Borghini, "Investigating driver fatigue versus alertness using the granger causality network," *Sensors*, vol. 15, no. 8, pp. 19181–19198, 2015.

[22] C. Zhao, M. Zhao, Y. Yang, j. Gao, N. Rao, and P. Lin, "The Reorganization of Human Brain Networks Modulated by Driving Mental Fatigue," *IEEE Journal of Biomedical and Health Informatics*, vol. 2016, pp. 1–1, 2016.

[23] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, Y.-H. Pan, and Y.-H. Wang, "Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1649–1657, 2012.

[24] U. R. Acharya, F. Molinari, S. V. Sree, S. Chattopadhyay, K.-H. Ng, and J. S. Suri, "Automated diagnosis of epileptic EEG using entropies," *Biomedical Signal Processing and Control*, vol. 7, no. 4, pp. 401–408, 2012.

[25] Y. Xiong, J. Gao, Y. Yang, X. Yu, and W. Huang, "Classifying driving fatigue based on combined entropy measure using EEG signals," *International Journal of Control and Automation*, vol. 9, no. 3, pp. 329–338, 2016.

[26] R. Chai, G. Naik, T. N. Nguyen et al., "Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system," *IEEE Journal of Biomedical and Health Informatics*, vol. 99, p. 1, 2016.

[27] C. Zhang, H. Wang, and R. Fu, "Automated detection of driver fatigue based on entropy and complexity measures," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 168–177, 2014.

[28] S. Kar, M. Bhagat, and A. Routray, "EEG signal analysis for the assessment and quantification of driver's fatigue," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 13, no. 5, pp. 297–306, 2010.

[29] K. A. Lee, G. Hicks, and G. Nino-Murcia, "Validity and reliability of a scale to assess fatigue," *Psychiatry Research*, vol. 36, no. 3, pp. 291–298, 1991.

[30] G. Borg, "Psychophysical scaling with applications in physical work and the perception of exertion," *Scandinavian Journal of Work, Environment & Health*, vol. 16, no. 1, pp. 55–58, 1990.

[31] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*, vol. 271, Springer Science & Business Media, 2012.

[32] M. Azarnoosh, A. Motie Nasrabadi, M. R. Mohammadi, and M. Firoozabadi, "Investigation of mental fatigue through EEG signal processing based on nonlinear analysis: symbolic dynamics," *Chaos, Solitons & Fractals*, vol. 44, no. 12, pp. 1054–1062, 2011.

[33] N. Kannathal, M. L. Choo, U. R. Acharya, and P. K. Sadasivan, "Entropies for detection of epilepsy in EEG," *Computer Methods and Programs in Biomedicine*, vol. 80, no. 3, pp. 187–194, 2005.

[34] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 6, pp. 2297–2301, 1991.

[35] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate and sample entropy," *American Journal of Physiology—Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[36] Y. Song, J. Crowcroft, and J. Zhang, "Automatic epileptic seizure detection in EEGs based on optimized sample entropy and extreme learning machine," *Journal of Neuroscience Methods*, vol. 210, no. 2, pp. 132–146, 2012.

[37] J. Xiang, C. Li, H. Li et al., "The detection of epileptic seizure signals based on fuzzy entropy," *Journal of Neuroscience Methods*, vol. 243, pp. 18–25, 2015.

[38] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou, "The appropriate use of approximate entropy and sample entropy with short data sets," *Annals of Biomedical Engineering*, vol. 41, no. 2, pp. 349–365, 2013.

[39] X. Li, B. Hu, S. Sun, and H. Cai, "EEG-based mild depressive detection using feature selection methods and classifiers," *Computer Methods & Programs in Biomedicine*, vol. 136, pp. 151–161, 2016.

[40] Z. Mu, J. Hu, and J. Min, "EEG-based person authentication using a fuzzy entropy-related approach with two electrodes," *Entropy*, vol. 18, no. 12, article 432, 2016.

[41] M. Zhong, F. Lotte, M. Girolami, and A. Lécuyer, "Classifying EEG for brain computer interfaces using Gaussian processes," *Pattern Recognition Letters*, vol. 29, no. 3, pp. 354–359, 2008.

[42] K. Polat and S. Güneş, "Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform," *Applied Mathematics & Computation*, vol. 187, no. 2, pp. 1017–1026, 2007.

[43] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier," *Computer Methods & Programs in Biomedicine*, vol. 108, no. 1, pp. 10–19, 2012.

[44] U. Orhan, M. Hekim, and M. Ozer, "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13475–13481, 2011.

[45] M. Sabeti, S. Katebi, and R. Boostani, "Entropy and complexity measures for EEG signal classification of schizophrenic and control participants," *Artificial Intelligence in Medicine*, vol. 47, no. 3, pp. 263–274, 2009.

[46] L.-N. Do, H.-J. Yang, S.-H. Kim, G.-S. Lee, and S.-H. Kim, "A multi-voxel-activity-based feature selection method for human cognitive states classification by functional magnetic resonance imaging data," *Cluster Computing*, vol. 18, no. 1, pp. 199–208, 2015.

[47] C. Vidaurre, A. Schlögl, R. Cabeza, R. Scherer, and G. Pfurtscheller, "Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 3, article no. 28, pp. 550–556, 2007.

[48] A. T. Azar and S. A. El-Said, "Performance analysis of support vector machines classifiers in breast cancer mammography recognition," *Neural Computing and Applications*, vol. 24, no. 5, pp. 1163–1177, 2014.

[49] J. Hu, Z. Mu, and P. Wang, "Multi-feature authentication system based on event evoked electroencephalogram," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 4, pp. 862–870, 2015.

[50] X. Bao, J. Wang, and J. Hu, "Method of individual identification based on electroencephalogram analysis," in *Proceedings of the International Conference on New Trends in Information and Service Science (NISS '09)*, pp. 390–393, Beijing, China, July 2009.

[51] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 121–131, 2011.

[52] X. Zhao, S. Xu, J. Rong, and X. Zhang, "Discriminating threshold of driving fatigue based on the electroencephalography sample entropy by receiver operating characteristic curve analysis," *Xinan Jiaotong Daxue Xuebao/Journal of Southwest Jiaotong University*, vol. 48, no. 1, pp. 178–183, 2013.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 2938504, 13 pages
https://doi.org/10.1155/2017/2938504

Research Article

Assessment of Iterative Closest Point Registration Accuracy for Different Phantom Surfaces Captured by an Optical 3D Sensor in Radiotherapy

Gerald Krell,[1] Nazila Saeid Nezhad,[1] Mathias Walke,[2] Ayoub Al-Hamadi,[1] and Günther Gademann[2]

[1]*Institute for Information Technology and Communication Engineering, Otto-von-Guericke University Magdeburg, Universitätsplatz 2, 39016 Magdeburg, Germany*
[2]*Clinic for Radiotherapy, Otto-von-Guericke University Magdeburg, Leipziger Straße 44, 39120 Magdeburg, Germany*

Correspondence should be addressed to Gerald Krell; krell@ovgu.de

Received 4 July 2016; Revised 30 September 2016; Accepted 25 October 2016; Published 9 January 2017

Academic Editor: Ayman El-Baz

An optical 3D sensor provides an additional tool for verification of correct patient settlement on a Tomotherapy treatment machine. The patient's position in the actual treatment is compared with the intended position defined in treatment planning. A commercially available optical 3D sensor measures parts of the body surface and estimates the deviation from the desired position without markers. The registration precision of the in-built algorithm and of selected ICP (iterative closest point) algorithms is investigated on surface data of specially designed phantoms captured by the optical 3D sensor for predefined shifts of the treatment table. A rigid body transform is compared with the actual displacement to check registration reliability for predefined limits. The curvature type of investigated phantom bodies has a strong influence on registration result which is more critical for surfaces of low curvature. We investigated the registration accuracy of the optical 3D sensor for the chosen phantoms and compared the results with selected unconstrained ICP algorithms. Safe registration within the clinical limits is only possible for uniquely shaped surface regions, but error metrics based on surface normals improve translational registration. Large registration errors clearly hint at setup deviations, whereas small values do not guarantee correct positioning.

1. Introduction

Tomotherapy combines a CT scanner with a computer-controlled radiation beam collimation system at the treatment machine [1] to precisely target tumors sparing healthy tissue. The system installed in Magdeburg hospital is a Tomotherapy HD system which enables helical and fixed radiation in one single system. A helical slit delivers radiation with most conformal image guided radiotherapy (imrt). The x-ray source rotates in a helical path around the patient in order to acquire a 3D image. The same x-ray source is used as treatment beam. This source is rotating in a helical pattern around the patient, while the intensity of beam is modulated according to the tumor shape using "tungsten leaves." These leaves create thousands of beam elements, called "beamlets" [2]. The radiation is delivered by a discrete-angle, nonrotational method sequentially moving the treatment table from the center of the system and for each angle of the gantry. Optical sensors provide an additional tool to verify the precise positioning of the radiation target relative to the treatment machine. The actual position in the treatment fraction is compared with the desired position given by a previously recorded reference surface. The reliability of such ICP-based algorithms is investigated in this paper by comparing the results of the implementation by the optical sensor with selected popular algorithms.

2. Methods and Materials

2.1. Optical Sensors in Tomotherapy. Nowadays, image guided methods are increasingly used in radiotherapy [3–11]. The

target regions of irradiation and the intended dose distributions are mostly defined on the basis of CT scans. Then an irradiation plan is created which involves the placement of the patient with regard to the treatment machine and the control of the irradiation beam. The main aim is to hit the tumor with sufficient energy and to protect healthy tissue and organs as much as possible against irradiation at the same time. Exact placement of the patient in the irradiation session is therefore very important. In addition to correct positioning by integrated CT in the treatment machine, optical sensors can capture surface data.

Optical surface sensors hence provide an additional tool for contact-less verification of patient position and are now getting into clinical practice after a long-time development and use for scientific purposes.

Our Tomotherapy HD accelerator unit (Accuray, USA) is combined with an AlignRT (VisionRT Ltd., London, UK) system and consists of two pods laterally positioned corresponding to the virtual isocenter in front of the Tomotherapy bore. The virtual isocenter lies 700 mm outside in front of to the real radiation isocenter of the machine. The distances of the two pods in respect to the virtual isocenter are about 2.0 m. The two pods are tightly mounted at upper ceiling of the treatment room. They are right and left of the Tomotherapy couch. Each of the two units each consists of two cameras (stereo) and a speckle projector producing structured light (Figure 1(a)) to generate a 3D model of the patient's surface by close-range photogrammetry (triangulation) [12]. The unit also includes a texture camera for visualization purposes, which, however, is not used for alignment [3]. The AlignRT parameters of the optical system are estimated and verified by daily calibration using a calibration plate that is aligned with a virtual laser isocenter in front of the real isocenter. Real-time capability of the AlignRT system relates to the ability of the sensor to capture surface data fast enough to even follow typical human motion caused by respiration, for example. Although tracking of the surface is fast enough to meet these requirements, the first registration takes longer and is therefore usually done offline.

The units are installed at the ceiling in the treatment room above of the treatment table and in front of the irradiation gantry in such a way that they capture the body surface at the target region (isocenter) diagonally downwards from two directions in order to reduce occluded regions (Figure 1(b)). The radiation gantry of the treatment machine is situated on one point of a circle around the isocenter parallel to the xz-axes. The x, y, and z position of the treatment table at the radiation gantry can be shifted computer-controlled with an accuracy of about 0.5 mm. Rotation of the treatment table is not possible, although in real situations rotational displacement of the patient must be expected.

An optical sensor of the considered type estimates a distance map related to a measured surface by finding correspondences in images taken from two or more directions by photogrammetric methods [12, 13]. A typical scheme is first to calculate a standard view of the recorded images by rectifying them on the basis of the camera parameters obtained by the previous calibration. Finding the correspondences in the images gives the disparity maps which describe the parallax caused by the distance between the cameras of one optical sensor. Together with camera calibration parameters the depth map is then calculated from the disparity map which can be considered as a mesh of 3D points or as a point cloud. Because the depth values are calculated corresponding to the pixel grid neighborhood relations are directly given and a mesh grid, for instance, consisting of triangles, can be easily calculated. The surfaces of the two optical sensors of AlignRT are merged in one data file. At the transition of the surface data of one sensor to the other some overlapping or gaps may occur. The software of the optical sensor handles these problems and produces a single more or less closed surface of triangles out of the data of the two sensors. Details are not given by the manufacturer. Rigid registration parameters for different snapshots of captured 3D surface data can be calculated by the propriety software.

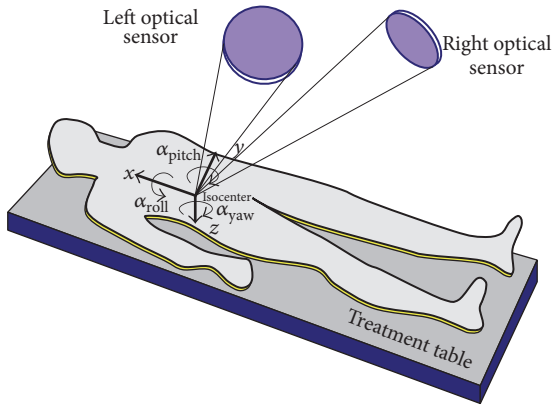
Optical sensors provide an additional modality to estimate the patient position on the basis of the outer body shape without increasing radiation load. Here we consider the application of the optical sensor without use of additional markers. The surface data captured is therefore a point cloud or a mesh grid corresponding to pixels of the image sensor. In such an unconstrained setting without markers, we just know that a surface point estimated by the optical sensor belongs to some corresponding point of the surface in the voxel image captured for definition of the 3D planning volume. But the exact position of this corresponding point on the surface in this image, which is usually a CT scan, is not directly given. This correspondence can only be estimated out of the form of the reference surface if it is successfully matched with the surface to be tested. In this way corresponding regions are registered and transformation matrices are calculated representing a measure for the deviation between a reference and a test surface.

Two operation modes of the surface sensor are distinguished in clinical practice: the static setup verification of patient (single-frame surface acquisition) and the tracking of patient motion (continuous dynamic surface acquisition mode), for example, caused by respiration [3]. In the latter case the solid assumption is appropriate if the time step from one surface capture to the next is small enough.

For real patients, the shape may also have been changed in the static setup phase introducing additional uncertainties in ICP registration. Respiratory motion of a patient surface blurs the registration result which is an additional effect and should not be mixed with the uncertainty of ICP registration. Motion blur in the 3D measurement results may cause additional problems. Our paper therefore considers the static case and shows that even with solid phantoms uncertainties remain depending on the individual shape.

2.2. Principles of ICP Algorithms. Surface registration assumes that two or more surfaces can be matched by a geometrical transformation. Resulting transformation parameters then describe the deviation between the surfaces for a correct registration. In case of radiotherapy, the optical sensor should ensure that the patient is placed according to the irradiation plan. The desired position is usually defined on basis of the CT data set. If we want to compare a test surface measured

(a) One of two camera-projector units of AlignRT (optical sensor) at the ceiling in treatment room. It consists of a stereo camera and a projector producing structured light by laser speckle in a fixed arrangement



(b) Left and right optical sensors at the ceiling above of the treatment table "look" diagonally downwards at the region of irradiation (isocenter). Standing in front of the treatment machine, the y-axis of the right-handed coordinate system points to the treatment machine, x to the right and z to the top. The irradiation gantry hitting the same region is not depicted. α_{roll}, α_{pitch}, and α_{yaw} describe the rotations around the x-, y-, and z-axes, respectively, as shown. The treatment table can only be translationally shifted in x, y, and z direction

Figure 1: Optical sensor in radiotherapy.

by optical sensors with the target position we would have to extract the corresponding surface data out of the CT as a reference. Tomotherapy gives us another option: because a CT is directly available at the treatment machine together with the optical sensor, we are able to bring the patient exactly to the desired position by the CT modality. The surface scan of the optical sensor at this position can be considered as a reference (target position). In later treatment sessions, the memorized reference position can then be used to bring the patient back to the desired position by the measured test surface. Also during the irradiation itself the correct position can be verified by the optical surface scan because, in contrast to CT, optical data are available during the whole treatment.

2.2.1. Known Point-to-Point Correspondences. The alignment of surfaces is much more simple and unique in case of known correspondences between reference and test surface. This assumes that registered n points of the reference surface $p_i = p(x_i, y_i, z_i)$ with index $i = 1, \ldots, n$ and spatial coordinates x_i, y_i, z_i are ordered in pairs with n points $q_i = q(x_i, y_i, z_i)$ of the test surface.

A linear transformation matrix R and a translation offset vector t aligning reference and test surface are directly estimated by minimization of the sum of the squared error:

$$E(R,t) = \sum_{i=1}^{n} \left\| p_i - Rq_i - t \right\|^2 \longrightarrow \text{Min}, \tag{1}$$

which means that the (geometric) distance between the reference surface and the transformed test surface should be as small as possible. When the correct correspondences are known a unique solution for R and t for given $n = N$ point pairs or a solution in the least square sense for $n > N$ directly yields. When limiting to an affine transform, the linear transform matrix with N^2 parameters modifies to a rotation matrix with N rotation angles resulting in a $2N$-dimensional optimization problem (together with the N translation parameters). Such a nonlinear equation can generally only be solved iteratively or with a linearized approximation assuming small angles.

2.2.2. Unknown Point-to-Point Correspondences. In general, without fiducial markers, no direct correspondence between points of the surfaces is given and also the number of points to be registered may be different. In this situation, the transformation matrix cannot be determined immediately. In ICP algorithms, the closest point in the reference is considered as the corresponding point of the test surface iteratively adapting the transformation matrix in each iterative step. Sophisticated search strategies exist in order to avoid a complete search between the two surfaces. The transformation in each iterative step does not align the two surfaces perfectly but brings them closer to each other in the converging case.

Registration fails in the case of growing deviation between the two surfaces in the iterative steps. When converging, the registration is terminated by a certain criterion such as size or gradient of the deviation error; that is, a certain registration error generally remains for real measurement data. ICP algorithms perform a local search on the error surface describing the deviation of the actual measurement from the target and estimate translation and rotation matrices as registration parameter. They converge well when a unique error minimum exists, but problems may arise when trapping in side minima occurs. In the latter case, registration is inaccurate or fails completely.

Reference [14] gives a good overview on ICP algorithms for technical applications with three synthetically generated

scenes providing test surfaces to evaluate the variants. In this way the correct transform is known exactly.

ICP algorithms can be divided into different phases. According to [14], typical ICP algorithms perform the following steps.

(1) Selection of the Source Points (Measurement). Different criteria for handling point clouds are considered. Using the complete set of points to find the transformation parameters might be slow; therefore the data could be randomly or regularly subsampled. Another strategy is to extract significant points at edges or corners where the information is concentrated. This method of sampling requires preprocessing but it reduces the number of required points improving accuracy and efficiency of the algorithm.

(2) Matching. This step is the most costly step in ICP algorithm. There are different methods such as building a kd-tree search to speed up finding corresponding closest points. The simplest idea is finding the closest point in the other point cloud for each point. The result of this method is generally stable but it computes slowly. Another method to find the correspondence is "shooting" along the normal of each point to the other point cloud. The intersection of the normal and point cloud is considered as the corresponding point [14]. There is a faster method to match the correspondence which is projection based matching. In this method the points lying on the line of sight of one of the cameras are considered correspondent. In this case the result is good if two cameras are close enough [14].

(3) Weighting. The matched point pairs can be weighted with regard to certain additional criteria describing the similarity of the corresponding region such as color, distance, curvature, or direction of tangent normal [14]. To this end, the error metric is multiplied by a weighting factor depending on the specific criterion.

(4) Rejection. Rejection of certain point pairs can be implemented after each matching step in order to improve alignment. This can be done in the phase of search for the closest neighbor. Several rejection methods have been proposed in different studies [14]: rejection of those point pairs with a distance greater than a user specified limit, rejection of a certain portion of point pairs with largest distance, and rejection of point pairs inconsistent with neighbor pairs (rigid transform).

(5) Error Metric and Minimization. This step is the last step of ICP algorithm which measures the error between the point clouds and tries to minimize the distance between two point clouds. Mostly either a "point-to-point" or a "point-to-plane" error metric is applied. In the first case, if p_i is a source point and q_i the corresponding point in the target point cloud and M is the transformation matrix, then the sum of squared distances has to be calculated and minimized [2]:

$$M_{\min} = \arg\min_{M} \sum^{i} \left(M \cdot p_i - q_i\right)^2. \tag{2}$$

Closed form solutions for this kind of error metric exist, such as singular value decomposition (SVD), dual quaternions, quaternions, and orthonormal matrices. Accuracy and stability of these methods have been evaluated by [15].

In general, point-to-plane error metric converges better than the point-to-point error metric [16]. It minimizes the sum of squared distances between source points and the tangent plane at the target point which is orthogonal to the unit normal vector of that point. Mathematically, if p_i is a source point and q_i is the corresponding point in the target point cloud and $n_i = n(x_i, y_i, z_i)$ is the normal vector at q_i then the ICP algorithm estimates the rigid transformation matrix by the minimizing function

$$M_{\min}^{\text{norm}} = \arg\min_{M} \sum \left(\left(M \cdot p_i - q_i\right) \cdot n\right)^2. \tag{3}$$

Because no closed form solutions for point-to-plane error metric exist it is usually solved iteratively by nonlinear methods such as Levenberg-Marquardt or it can be linearized considering some approximation for rotation matrix R, such as replacing $\sin\theta$ by θ and $\cos\theta$ by 1. The problem of the point-to-plane error metric is that it is sensitive to noise and that it does not converge well if the distance between two point clouds is large [15, 17].

The ICP algorithm can vary by changing the methods in each step to improve the performance with regard to speed and stability depending on the amount of noise and outliers the algorithm can deal with.

2.3. Selected ICP Algorithms for Registration. Four different, under BSD license available, ICP implementations in Matlab have been compared with the proprietary software of AlignRT for surface registration of phantoms. We have chosen the same software platform because one criterion was the option to compare the speeds. We assumed that the implementations belong to the most popular ones. They all meet the same general ideas of ICP registration and present the variety of unconstrained methods (without markers or using colors). We found that the four chosen ICP algorithms are well suited to be compared with the method applied by AlignRT. An interesting extension of work would be to include new approaches to point registration such as described in [18].

(1) Wilm's Algorithm [2, 19]. Point clouds are aligned by considering the complete points set. The program finds the nearest neighbor by a kd-tree search which considerably increases the speed of matching. Point-to-point or point-to-plane error metric can be selected by parameter setting. AlignRT uses a similar point-to-plane metric as follows from the communication with the manufacturer.

(2) Kroon's [20], Modified. This program uses a finite difference model to align the point clouds. The finite difference method also supports the transform types of resizing and shearing. Several optimization functions are included for minimum search. We added a global search approach by generating different start points using a scatter-search method to improve the results. All starts points are evaluated and

the points which are unlikely to improve the minimum are rejected.

(3) Renoald's [21]. It is a simple ICP implementation which uses all the data points. It first finds the corresponding points by creating a Delaunay tessellation of points in a model to search for the closest point. Then it calculates the initial transformation matrix by singular value decomposition (SVD) and applies this to the target point cloud. The transformation matrix is updated iteratively until no more correspondences can be found.

(4) Bergström's [22, 23]. It is similar to the Renoald's algorithm with the main difference that, after matching corresponding points, the point pairs are weighted by the maximum point distance. Levenberg-Marquardt algorithm is directly applied to minimize the squared sum of the distances of closest points.

Most of the implementations allow choosing among modes and modifying parameters. The best configuration for this experimental setup has been investigated and shown in Table 3. The above given references give further details.

2.4. Related Works. Reference [4] compared suggested setup correction with a second and independently operated marker-based optical system with an anthropomorphic plastic phantom and healthy volunteers. They found alignment accuracies of about 1 mm for translation and 0.5° for rotation as an average. Using markers is more invasive and time consuming but in general safer than unconstrained registration.

Extensive research has been done on the development of surface sensors. The general ideas are shown in works as [12, 13]. Reference [24] deals with the simulation of photogrammetric triangulation in order to develop the algorithms without need of acquisition of additional camera data.

Reference [3] investigated the temporal stability of alignment accuracy in the context of respiratory motion in an operation mode where the sensor is triggered by the breathing phase. A rigid, flesh-colored mannequin torso phantom has been used. In this approach, the optical sensor is combined with an infrared-based marker system for gating the breathing state and a motorized mechanical stage. Measured surface data has been compared with surface extracted from CT as a reference. High stability and errors in the submillimeter range and less than 1° have been reported. Additionally, the accuracy of recommended patient realignment has been evaluated for 54 random shifts of the treatment table. In our investigations we focused our attention on the influence of different types of phantoms in order to learn how curvature influences the registration reliability.

Reference [14] gives helpful results how existing ICP algorithms converge for synthesized surfaces. Also different sampling strategies for selection of registration points have been considered. But for clinical practice it is important to verify these theoretical results with the real situation for data of an existing optical sensor.

Reference [7] evaluates a 3-Dimensional Surface Imaging System for Guidance in DIBH Therapy. Setup data based on captured 3D surfaces by the same surface imaging system as we used was compared with setup data based on cone beam computed tomography (CBCT) and evaluated with regression based methods. It was found that in the context of breast cancer treatment 95% of the deviations less than 0.4 cm detected by the optical sensor were less than 0.66 cm in the other mode of CBCT. A comparison of megavoltage CBCT based registration and of AlignRT based registration to its own particular reference is subjected to certain time constraints. A CT scan itself as a possible reference and the local megavoltage CBCT scan on the Tomotherapy unit is usually a time-consuming procedure.

Reference [6] reports on two commercial optical sensors (surface imaging systems) and compares them with the actual adjustments in patient positions made on the basis of megavoltage CT scans. The deviations between the proposed correction of the optical sensor and the subjectively best alignment of an expert have been statistically evaluated. Tests have been performed on an Alderson phantom and on patients at head/neck, pelvic, and chest regions. It was found that the optical sensors can support patient positioning mainly at pelvic and chest regions because immobilization of the patient by special masks is not possible as in the case of head and neck region.

Generally, the AlignRT system is usable on nearly all possible patient regions. Some papers deal with clinical applications of optical sensors to different patient regions. Besides classical patient body region dependent applications, the frame- and mask-less cranial stereotactic radiosurgery is a new application field. The comparison of breath induced surface movements with different registration modalities is subjected to different time constants of the acquisition devices. The verification of DIBH (depth inspiration breath-hold) techniques with optical systems, as the AlignRT system, is a new emerging procedure in the clinical practice.

A feasibility study for the usability of the AlignRT system to frame- and mask-less cranial stereotactic was presented by [8]. The presented technique shows the potential of head mold and surface monitoring to use in stereotactic treatments. The accuracy of the surface imaging motion tracking system during the stereotactic treatment was verified. The results were additionally tested on the standard optical guidance platform technique (kVCT by Varian).

Work [9] describes a clinical analysis of fifty patients with the AlignRT system in comparison to megavoltage portal imaging. Daily alignment with the 3D optical imaging system was found to be valuable for reducing setup errors in comparison to skin markers. Particularly the anterior-posterior alignment directions were with the optical system noticeably better.

The possible synchronization of a classical CBCT system with the AlignRT has been shown by [10]. An image guided method for the synchronization of the X-ray projections is synchronized with optically sensed surface during using CBCT without any further hardware requirements. The proposed method can by generically applied to any configuration of the CBCT and optical imaging systems and also be used for extracranial tumor tracking.

3. Generation of Test Surface Data

In order to generate surface data we focused our work on rigid phantoms because we are mostly interested in pure accuracy of the sensor together with the ICP algorithms in the ideal case. The investigated ICP algorithms do not treat shape variations which is a motivation for using solid phantoms instead of real cases. The influence of motion of real human bodies, caused by respiration for instance, is considered by other papers (e.g., [10, 11]).

3.1. Test Phantoms. Because the contour characteristics of a surface is important for a safe registration, specially designed phantoms of different surface types have been investigated. To this end, dedicated phantoms have been designed or selected with a size approximately covering the measurement volume of the optical sensor of about $0.1\,\mathrm{m}^3$. In this study, four different phantoms have been measured by the optical sensor in order to generate point clouds for the evaluation of the ICP algorithms.

(i) Plane. It is a simple plane horizontally placed on the treatment table. The main idea is to check the accuracy of the optical sensor with regard to vertical shift of the treatment table (z direction).

(ii) 3plane. It consists of two planes and an edge especially built to allow a unique matching with respect to all x-y-z space coordinates.

(iii) Bowl. The bowl phantom is more curved than a plane, but ambiguities with regard to rotations must be expected.

(iv) Torso. By the torso of a mannequin, a shape typical for the human body has been simulated. The curvature of the torso phantom is more ambiguous in the cranial-caudal (y) than in the dorsal-ventral transverse motion direction.

The phantoms have been coated by white painting or textile to produce a surface that can be well captured by the cameras of the optical sensor when illuminated by the speckle projector. Measured point clouds of these phantoms are shown in Figure 2. As visible in Figure 2(c), the measured surfaces contained some points of the background (e.g., of the treatment table). Such extra points obviously not originating from the phantoms have been manually removed for the data of all phantoms. As an example, Figures 2(c) and 2(d) show the bowl surface before and after removing the extra points, respectively.

3.2. Test Setup. The above described ICP algorithms have been tested with surface data of the selected phantoms (Figure 2) moved to well-defined positions. First, the optical sensor AlignRT has been calibrated according to the instructions of the manufacturer. Then the phantoms have been placed on the treatment table and a surface scan at the origin has been captured. This surface scan at central (zero) position of the treatment table served as a reference to compare with surface scans at other positions. To this end, the phantoms have been translationally shifted by the treatment table in the directions $d = \{x, y, z\}$ by distances of $s_d = \{0.5, 1.0, \pm 10.0, 20.0\}$ (mm). For the plane, translation was only done in z direction ($d = z$) because tests confirm the obvious fact that a motion in x or y direction cannot be detected if the plane phantom is placed in parallel to the x-y-axes as we did.

Figure 3 gives an example on how the operator sees the situation on the monitoring screen of AlignRT. It shows the estimated misalignment for translation and rotation in mm and °, respectively, by numbers with one-digit accuracy after the comma and by bars. At setup, the therapists attempt to minimize the shifts (by minimizing the length of the bars) [11]. The surface data is exported as object files and used for the registration by the other ICP algorithms.

After an initial phase, real-time surface tracking is possible with the AlignRT system. AlignRT system delivers sufficiently fast displacement estimation for most medical indications of about 10 frames per second. Acceptable speed relates mainly to the time needed for an initial alignment which should not exceed about a second in order to be acceptable in clinical routine.

Therefore two requirements result with regard to the speed: the alignment time should not be much longer than a second because more cannot be accepted in clinical routine.

In case of dynamic tracking, the speed demands arise by the typical patient motion to avoid subsampling on the one hand and to ensure that shape variations between two time steps can be neglected for rigid registration. In the ideal case, the registration should be faster than the surface sensor in order to avoid reduction of frame rate.

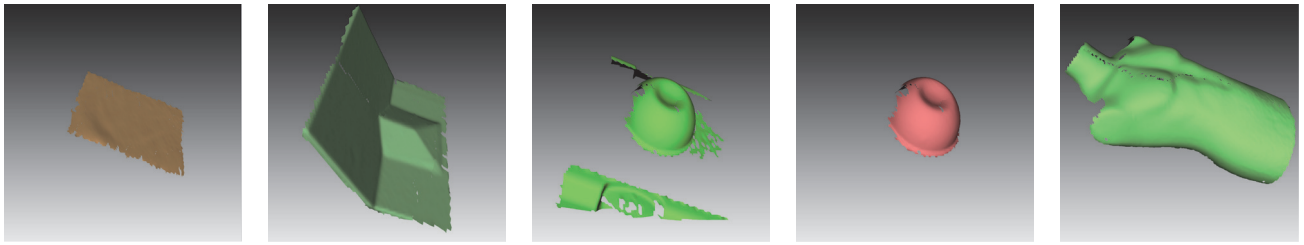
4. Results and Discussion

Rigid transform matrices (translation and rotation) for registration of the reference with the tested position have been estimated by the proposed ICP algorithms and with AlignRT. The investigated implementations specify the resulting coordinate transform for registration by different versions of matrices for homogenous coordinates. For direct comparison, these matrices have been transformed into a single representation for translation and rotation (see [12]).

Translational shift values of registration $\widehat{s}_d$ in direction d yield directly from the offset part of the transform matrices. Table 1 shows the results of registration together with the expected translation values s_d. The translational registration error in direction d is then given by $e_d^{\mathrm{trans}} = s_d - \widehat{s}_d$.

The total registered rotation is composed by a series of three rotations $r = \{\mathrm{roll}, \mathrm{pitch}, \mathrm{yaw}\}$ around x-, y-, and z-axes, respectively, in the directions according to Figure 1(b), each quantified by the Euler angles $\widehat{\alpha}_r = \{\widehat{\alpha}_{\mathrm{roll}}, \widehat{\alpha}_{\mathrm{pitch}}, \widehat{\alpha}_{\mathrm{yaw}}\}$. The rotatory registration error is $e_r^{\mathrm{rot}} = \widehat{\alpha}_r - \alpha_r = \widehat{\alpha}_r$ for $\alpha_r = 0$ because the measurement phantoms have not been rotated.

We assumed a maximally allowed absolute registration error of $e_{d\,\mathrm{max}}^{\mathrm{trans}} = 1\,\mathrm{mm}$ for translation and of $e_{d\,\mathrm{max}}^{\mathrm{rot}} = 0.5^\circ$ for rotation which are quite tough values in radiotherapy and marked entries with $|e_d^{\mathrm{trans}}| > e_{d\,\mathrm{max}}^{\mathrm{trans}}$ or $|e_d^{\mathrm{trans}}| > e_{d\,\mathrm{max}}^{\mathrm{rot}}$ boldface. Other works set the allowable tolerance a bit higher



(a) Plane (b) 3plane (c) Bowl (before removing extra points) (d) Bowl (after removing the extra points) (e) Torso

Figure 2: Surfaces of selected phantoms captured by the optical sensor AlginRT showing typical problems of real measurement data: (a) measurement noise and systematic errors; (c) extra points not belonging to the object of interest; (b) and (e) seam from fusing the two surfaces of left and right optical sensors.

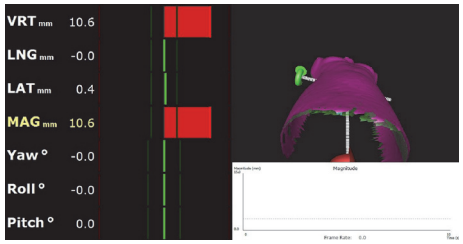


Figure 3: An example of the AlignRT monitoring screen seen during measurement of the torso phantom and vertical shift of the treatment table $s_d(z) = 10$ mm. The reference surface is shown in pink and the measured surface in green. The suggested linear translations (vertical, lateral, and longitudinal) and rotations (yaw, pitch, and roll) are shown by numbers and colored bars on the left together with the RMS value (called magnitude MAG). The white graph is used to display a time series of the RMS values (not used in our experiment).

(e.g., [8] to 1 mm/1° and [11] to 3 mm/3°), but working with rigid phantoms without motion motivates our stricter limits.

Figure 4 shows as an example one of the best results of aligned surfaces with the reference surface for a shift of $s_d = 10$ mm in direction $d = z$ using the Wilm approach. The residuals have been estimated by triangulation of the surfaces and color-coded displaying the distances in z direction. It becomes clear that although the translational and rotational parameters are within the limits this does not hold for all points of the surfaces. There are problems especially at sloping surface parts, at edges, and at the stitching area of left and right optical sensors which explains the remaining deviations after applying the ICP algorithm.

Table 2 summarizes Table 1 with regard to adhering the limits $e_{d\,\max}^{\text{trans}}$ and $e_{d\,\max}^{\text{rot}}$. As expected with the plane phantom placed in parallel to the xy plane a safe registration is only possible in z direction and fails in x and y direction for all ICP algorithms. For the other three phantoms 3plane, bowl, and torso, only the Wilm algorithm registers safely for the translational parameters. No algorithm has problems with the rotatory parameters for any phantom except Wilm which interestingly fails for the torso phantom for pitch and yaw and AlignRT for yaw of the bowl phantom.

Table 3 compares some important properties and results of the four tested algorithms that have been applied to four different test objects (phantoms) differently shifted relative to an original position. The algorithms use different methods to compute the rigid transformation matrix (translation and rotation) between two point clouds, as described in Section 2.3 as the result of registration.

Main operational principles of the algorithms are summarized; their processing speed and accuracy give information on their suitability for registration of our selected phantoms. Main differences consist in the method for the closest point search, the weighting, the error metric, and the method for minimization. Only Wilm uses kd-tree search which is much more efficient than full search. Only Bergström applies distance-based weighting. None of the open source algorithms includes rejection. Among the open source algorithms, only Wilm uses point-to-plane metric whereas all other apply a point-to-point criterion. The AlignRT registration results look similar to the Wilm implementation. This supports the assumption that similar principles are used by this proprietary program.

The average processing time for each algorithm is also qualitatively given. It varies between fastest processing (which was about a few seconds) and slowest processing (which was about 3 minutes) for the registration by the ICP algorithm on a standard computer (Intel Core i7, 64-bit Windows) in Matlab. A more detailed evaluation of processing speed is not given because we do not expect that the chosen algorithms are implemented in an optimal way. This may be different for the commercial implementation of AlignRT. Renoald performed best with regard to processing speed. Wilm and AlignRT show acceptable speed in the same range. Kroon is slow and Bergström is very slow in the investigated implementation and would not be acceptable in clinical routine.

For offline verification, speed plays a less important role as long as the registration takes only seconds of time. Therefore those implementations indicated by + or ++ can be considered acceptable in the intended application (see Section 3.2). In tracking applications, when even the registration is done online, the speed of the algorithms matters much more and the patient alignment can be verified and corrected on the fly by moving the treatment table or adapting the irradiation. But

Table 1: Registered translations $\widehat{s}_x$, $\widehat{s}_y$, and $\widehat{s}_z$ and rotations $\widehat{\alpha}_{\text{roll}}$, $\widehat{\alpha}_{\text{pitch}}$, and $\widehat{\alpha}_{\text{yaw}}$ from the investigated sample implementations of the ICP algorithm for the tested phantoms plane, 3plane, bowl, and torso and treatment table shifts s_d in different directions $d = \{x, y, z\}$. Translations with an absolute translational registration error $|e_d^{\text{trans}}| > e_{d\ \max}^{\text{trans}} = 1$ mm and rotations with an absolute registration error $e_r^{\text{rot}} > e_{d\ \max}^{\text{rot}} = 0.5^\circ$ are marked boldface.

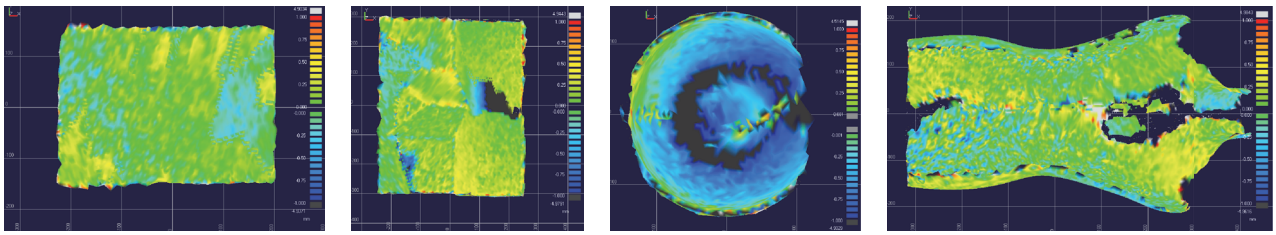
Phantom	Shift		Registration					
			Translation/mm			Rotation/°		
	s_d/mm	d	$\widehat{s}_x$	$\widehat{s}_y$	$\widehat{s}_z$	$\widehat{\alpha}_{\text{roll}}$	$\widehat{\alpha}_{\text{pitch}}$	$\widehat{\alpha}_{\text{yaw}}$
Wilm algorithm								
Plane	0.5	z	0.4282	0.0087	0.5482	−0.0001	−0.0012	0.0000
	1.0	z	**1.4503**	0.4656	1.0627	0.0000	−0.0123	0.0000
	10.0	z	0.4760	0.8346	10.0927	0.0002	−0.0022	−0.0005
	20.0	z	0.0827	**1.1255**	20.0975	−0.0001	0.0213	−0.0007
3plane	0.5	z	−0.0100	0.1440	0.5325	−0.0000	0.0001	−0.0003
	1.0	z	−0.0395	0.1362	1.0882	0.0001	0.0004	0.0001
	10.0	z	−0.1475	0.1217	10.0550	0.0003	0.0001	−0.0003
	20.0	z	−0.0665	−0.0903	20.9796	0.0005	0.0012	−0.0006
	10.0	y	−0.1237	9.9743	0.0157	0.0001	0.0004	0.0001
	−10.0	y	−0.1991	−9.6955	0.1602	0.0002	−0.0001	0.0004
	10.0	x	9.8886	0.3100	0.1109	0.0013	−0.0004	0.0001
Bowl	0.5	z	0.0749	0.0702	0.4491	−0.0006	0.0013	−0.0000
	1.0	z	0.0596	0.1477	0.9348	−0.0010	−0.0018	−0.0011
	10.0	z	−0.0084	0.0054	10.1927	0.0061	0.0121	0.0004
	20.0	z	0.3618	0.0843	20.0056	0.0005	0.0382	−0.0001
	10.0	y	−0.1668	9.9264	0.1325	0.0035	−0.0678	−0.0033
	−10.0	y	0.0885	−9.7892	−0.3034	0.0004	0.0246	−0.0006
	10.0	x	9.9217	−0.0522	0.0799	−0.0014	−0.0400	−0.0024
	−10.0	x	−9.8553	0.1800	−0.0533	−0.0005	0.0854	−0.0005
Torso	0.5	z	−0.0239	−0.0895	0.5171	0.0033	**1.5783**	**−1.5785**
	1.0	z	−0.0986	−0.1490	1.0508	0.0022	**1.6005**	**−1.6007**
	10.0	z	0.0303	0.1723	9.9136	0.0090	**1.1490**	**−1.1480**
	20.0	z	0.0995	0.1012	19.8599	0.0017	**0.8377**	**−0.8362**
	10.0	y	0.1819	9.9194	−0.1830	0.0009	**0.9571**	**0.9572**
	−10.0	y	−0.1553	−9.5344	0.0316	0.0020	**−1.4725**	**1.4724**
	10.0	x	10.1230	−0.0174	−0.0374	0.0047	**1.5300**	**−1.5299**
	−10.0	x	−9.8724	−0.0488	−0.0401	0.0034	**−1.6261**	**1.6263**
Kroon algorithm								
Plane	0.5	z	−0.4031	−0.1523	0.5499	−0.0001	−0.0001	−0.000
	1.0	z	−0.9254	−0.2737	1.0675	0.0002	0.0005	−0.0001
	10.0	z	1.3403	**1.2951**	10.0872	0.0002	−0.0015	−0.0004
	20.0	z	**−3.7160**	0.7362	20.0891	−0.0001	0.0007	−0.0008
3plane	0.5	z	0.0023	−0.1280	0.3746	−0.0005	0.0000	0.0000
	1.0	z	−0.0470	−0.2840	0.8252	−0.0001	0.0003	0.0001
	10.0	z	−1.5692	−5.8559	**7.3710**	−0.0125	−0.0045	−0.0043
	20.0	z	**1.0803**	**−4.9535**	**18.1806**	1.0803	0.0014	−0.0021
	10.0	y	0.0793	**6.5978**	**−1.3214**	−0.0066	−0.0004	−0.0003
	−10.0	y	−0.1597	**−6.6084**	**−1.3826**	0.0062	0.0004	0.0006
	10.0	x	**0.7600**	−0.0439	0.0272	0.0002	0.0008	−0.0018
Bowl	0.5	z	0.0818	−0.0440	0.2600	0.0009	0.0004	−0.0004
	1.0	z	0.0572	−0.0864	0.5162	0.0015	0.0002	−0.0004
	10.0	z	**−1.4703**	−0.2302	**8.9576**	0.0087	0.0022	−0.0420
	20.0	z	**−1.2950**	−0.1217	19.0383	0.00206	0.0035	−0.0554
	10.0	y	−0.2905	**8.8071**	0.0682	0.0369	−0.0581	−0.0027
	−10.0	y	0.2614	**−8.6043**	0.3671	−0.0368	0.0533	0.0006
	10.0	x	9.4381	−0.4111	0.0864	0.0023	0.0202	−0.0048
	−10.0	x	−9.1155	0.5139	0.1217	0.0004	−0.0187	0.0473

Table 1: Continued.

Phantom	Shift		Registration					
			Translation/mm			Rotation/°		
	s_d/mm	d	$\hat{s}_x$	$\hat{s}_y$	$\hat{s}_z$	$\hat{\alpha}_{\text{roll}}$	$\hat{\alpha}_{\text{pitch}}$	$\hat{\alpha}_{\text{yaw}}$
Torso	0.5	z	0.3474	**1.4598**	0.1776	−0.0003	0.0003	−0.0128
	1.0	z	0.3474	**1.5383**	0.0214	−0.0012	0.0013	−0.0214
	10.0	z	0.0014	0.2854	10.4856	0.0004	0.0018	−0.0006
	20.0	z	0.2456	0.8967	20.8858	0.0016	0.0010	−0.0012
	10.0	y	0.1391	**0.8154**	0.5305	0.0032	−0.0002	−0.0007
	−10.0	y	0.0276	−0.7041	−0.3240	−0.0025	0.0002	−0.0003
	10.0	x	10.7852	0.6038	0.0172	0.0002	0.0026	−0.0353
	−10.0	x	−10.6538	**−1.5451**	0.2848	0.0005	−0.0080	0.0327
Renoald algorithm								
Plane	0.5	z	−0.3951	−0.1486	0.5499	−0.0001	−0.0001	−0.0000
	1.0	z	−0.8802	−0.2661	1.0675	0.0000	0.0006	−0.0001
	10.0	z	**1.2549**	**1.1879**	10.0882	0.0002	−0.0010	−0.0005
	20.0	z	**−2.0547**	−0.4494	20.0903	−0.0001	−0.0040	−0.0006
3plane	0.5	z	0.0041	−0.1311	0.3722	−0.0005	0.0000	0.0000
	1.0	z	−0.0469	−0.2864	0.8242	−0.0006	0.0003	0.0001
	10.0	z	−0.0744	**−4.5305**	**8.2095**	−0.0099	−0.0003	−0.0039
	20.0	z	0.9315	**−6.7487**	**16.8058**	−0.0134	0.0058	0.0033
	10.0	y	0.1794	**4.0410**	**−1.1857**	−0.0124	−0.0004	−0.0002
	−10.0	y	−0.2608	**−5.4906**	**1.3925**	0.0088	0.0009	0.0005
	10.0	x	**0.7315**	−0.0448	0.0321	0.0002	0.0009	−0.0018
Bowl	0.5	z	0.0818	−0.0440	0.2600	0.0009	0.0004	−0.0004
	1.0	z	0.0571	−0.0831	0.5143	0.0015	0.0003	−0.0004
	10.0	z	**−1.2139**	−0.1348	**8.8075**	0.0069	0.0014	−0.0393
	20.0	z	**−1.5751**	−0.1130	**18.7352**	0.0056	0.0045	−0.0556
	10.0	y	−0.4060	**7.5070**	−0.0636	0.0356	−0.0402	−0.0014
	−10.0	y	0.3518	**−8.5498**	0.2583	−0.0358	0.0521	0.0009
	10.0	x	**8.8252**	−0.2963	−0.1059	−0.0000	0.0147	−0.0470
	−10.0	x	**−8.9940**	0.4230	0.1382	0.0007	−0.0171	0.0477
Torso	0.5	z	0.2920	0.4220	0.7578	0.0005	0.0012	−0.0321
	1.0	z	0.2378	0.4817	1.0779	0.0003	0.0009	−0.0315
	10.0	z	−0.0037	0.2426	10.4506	0.0001	0.0018	−0.0006
	20.0	z	0.3008	−0.1655	**18.7455**	−0.0011	0.0022	−0.0001
	10.0	y	0.1177	**0.7804**	0.4367	0.0028	−0.0002	−0.0004
	−10.0	y	0.0045	−0.6522	−0.2843	−0.0022	0.0001	−0.0001
	10.0	x	10.4139	0.4050	0.0860	0.0004	0.0015	−0.0327
	−10.0	x	−10.4768	−0.6117	0.1427	−0.0003	−0.0019	0.0334
Bergström algorithm								
Plane	0.5	z	0.0818	−0.0440	0.2600	0.0009	0.0004	−0.0003
	1.0	z	−0.9302	−0.2768	1.0675	0.0000	0.0005	−0.0001
	10.0	z	**−4.8742**	−0.3023	10.0866	0.0002	0.0013	−0.0008
	20.0	z	**−3.7203**	0.7381	20.0894	−0.0000	0.0001	−0.0008
3plane	0.5	z	0.0023	−0.12880	0.3746	−0.0005	0.0003	0.0000
	1.0	z	−0.0470	−0.2840	−0.8252	−0.0060	0.0003	0.0001
	10.0	z	**−8.9059**	**4.1357**	**11.6053**	0.0083	−0.0080	−0.0024
	20.0	z	**1.0816**	**−4.9472**	**18.1816**	−0.0095	0.0014	−0.0021
	10.0	y	**−2.1412**	**13.7773**	**1.6485**	0.0079	0.0060	0.0009
	−10.0	y	−0.1596	**−6.6095**	**1.3820**	0.0062	0.0004	0.0006
	10.0	x	**8.4082**	**2.8181**	**1.2289**	0.0064	−0.0053	−0.0034

Table 1: Continued.

Phantom	Shift		Registration					
			Translation/mm			Rotation/°		
	s_d/mm	d	$\widehat{s}_x$	$\widehat{s}_y$	$\widehat{s}_z$	$\widehat{\alpha}_{\text{roll}}$	$\widehat{\alpha}_{\text{pitch}}$	$\widehat{\alpha}_{\text{yaw}}$
Bowl	0.5	z	0.0818	−0.0440	0.2600	0.0009	0.0004	−0.0004
	1.0	z	0.0527	−0.0803	0.5211	0.0016	0.0003	−0.0004
	10.0	z	−0.1868	−0.0644	10.2913	−0.0002	−0.1752	0.0064
	20.0	z	0.6199	−0.0374	20.2321	0.0092	−0.1716	0.0172
	10.0	y	0.0319	10.0246	−0.4303	0.0058	−0.1689	0.0073
	−10.0	y	0.4347	−9.4345	0.9152	−0.0312	−0.1553	0.0238
	10.0	x	9.9406	**−2.1170**	0.6590	0.0190	0.0004	0.0001
	−10.0	x	−10.3737	**−1.6094**	0.3040	0.0133	0.0020	−0.0003
Torso	0.5	z	0.3383	**−2.5946**	0.8076	0.0004	0.0001	−0.0128
	1.0	z	0.1906	**−2.6231**	0.9682	−0.0000	0.0000	−0.0121
	10.0	z	0.0014	0.2854	10.4856	0.0040	0.0018	−0.0006
	20.0	z	−0.4539	−0.3521	19.4238	0.0008	0.0023	0.0150
	10.0	y	0.5342	**8.9866**	0.3959	0.0007	−0.0020	−0.0169
	−10.0	y	−0.1569	−8.6553	1.2477	−0.0070	0.0020	0.0012
	10.0	x	10.4504	**−2.5372**	0.4949	0.0003	0.0003	−0.0131
	−10.0	x	−10.2027	**2.5363**	0.3930	−0.0003	−0.0002	0.0120
AlignRT algorithm								
Plane	0.5	z	0.1	0.3	0.0	0.0	0.0	0.0
	1.0	z	0.2	0.1	0.5	0.0	0.0	0.1
	10.0	z	0.3	0.0	10.1	0.0	0.0	0.1
	20.0	z	0.2	**1.2**	20.1	0.0	0.0	0.2
3plane	0.5	z	0.0	0.1	0.5	0.0	0.0	0.0
	1.0	z	0.1	0.1	1.1	0.0	0.0	0.0
	10.0	z	0.0	0.2	10.2	0.0	0.0	0.0
	20.0	z	0.0	0.1	20.1	0.0	0.0	0.1
	10.0	y	0.2	10.2	0.2	0.0	0.0	0.0
	−10.0	y	0.1	−9.8	0.1	0.0	0.0	0.0
	10.0	x	9.8	0.2	0.1	0.0	0.0	0.0
Bowl	0.5	z	0.1	0.1	0.5	0.0	0.0	0.0
	1.0	z	0.1	0.2	1.1	0.0	0.0	0.0
	10.0	z	0.1	0.2	10.0	0.1	0.0	**1.2**
	20.0	z	0.3	0.3	19.9	0.1	0.0	**1.6**
	10.0	y	0.2	10.3	0.2	0.0	0.1	**1.6**
	−10.0	y	0.7	−9.8	0.2	0.1	0.1	**2.7**
	10.0	x	10.1	0.2	0.2	0.0	0.1	0.1
	−10.0	x	−10.0	0.5	0.2	0.0	0.0	**1.7**
Torso	0.5	z	0.1	0.0	0.6	0.0	0.0	0.0
	1.0	z	0.1	0.0	1.1	0.0	0.0	0.0
	10.0	z	0.1	0.0	9.9	0.1	0.0	0.1
	20.0	z	0.2	0.1	19.8	0.0	0.1	0.1
	10.0	y	0.2	10.1	0.2	0.0	0.0	0.0
	−10.0	y	0.2	−9.8	0.1	0.1	0.0	0.0
	10.0	x	10.0	0.2	0.1	0.0	0.0	0.0
	−10.0	x	−9.9	0.1	0.0	0.0	0.0	0.0



(a) Plane: all residuals are below $e_{d\ \max}^{\text{trans}}$

(b) 3plane: problems at edges and tilted planes

(c) Bowl: problems at sloping surface parts

(d) Torso: holes at stitching area of the two sensors

FIGURE 4: Residuals of surface pairs shifted by $s_d = 10$ mm in direction $d = z$ of the studied phantoms captured by the optical sensor AlginRT and aligned by the Wilm approach.

TABLE 2: Summary of registration success and fail for translations (x, y, z) and rotation (rot); + denotes success when the specified misalignment threshold is deceeded and, otherwise, – labels the failing when the threshold is exceeded.

		Algorithm																			
		Wilm				Kroon				Renoald				Bergström				AlignRT			
Motion		x	y	z	rot	x	y	z	rot	x	y	z	rot	x	y	z	rot	x	y	z	rot
Phantom	Plane	–	–	+	+	–	–	+	+	–	–	+	+	–	–	+	+	+	–	+	+
	3plane	+	+	+	+	–	–	–	+	–	–	–	+	–	–	–	+	+	+	+	+
	Bowl	+	+	+	+	–	–	–	+	–	–	–	+	+	–	+	+	+	+	+	–
	Torso	+	+	+	–	+	–	+	+	+	–	+	+	+	–	+	+	+	+	+	+

+: $|e_d^{\text{trans}}| < e_{d\ \max}^{\text{trans}}$ or $|e_r^{\text{rot}}| < e_{d\ \max}^{\text{rot}}$; –: $|e_d^{\text{trans}}| \geq e_{d\ \max}^{\text{trans}}$ or $|e_r^{\text{rot}}| \geq e_{d\ \max}^{\text{rot}}$.

TABLE 3: Overall assessment of the tested ICP algorithms.

Property	Algorithm				
	Wilm	Kroon	Renoald	Bergström	AlignRT
Closest point search	kd-tree	Full	Full	Full	—[a]
Weighting	None	None	None	Distance-based	—[a]
Rejection	None	None	None	None	—[a]
Error metric	Point-to-plane	Point-to-point	Point-to-point	Point-to-point	Point-to-plane
Minimization	Linearization of rotation matrix	Global search	SVD	Levenberg-Marquardt	—[a]
Speed[b]	+	–	++	––	+
Max. $\lvert e_r^{\text{rot}}\rvert$	<1.0 mm	>1.0 mm	>1.0 mm	>1.0 mm	<1.0 mm
Max. $\lvert e_d^{\text{trans}}\rvert$	>0.5°	<0.5°	<0.5°	<0.5°	>0.5°

[a] Unknown.
[b] ++: very fast; +: fast; –: slow; ––: very slow.

in this case, the algorithm needs much less iterations because the position differences from time step to time step are much smaller compared to the first alignment in the static case.

In Table 3, an overall assessment of the expected registration error between expected shift and the translation calculated by the ICP algorithms is given. Translation in x and y direction was omitted for the plane phantom because no registration was possible due to the missing structure in viewing field and therefore only the translation in the z direction is specified.

One observation from the experiments is that the distance of shift does not affect the registration accuracy much. Also the required time for convergence is not really affected, obviously because the algorithms adapt their step size according to the gradient.

Much more important are the structure and curvature of the surfaces to be aligned. With an ambiguous surface the error surface has flat areas where ICP algorithms are likely to stick in a local minimum. Registration fails in this case to align the surfaces [25].

Wilm's implementation shows the best results among the studied ICP algorithms for translational registration. The reason for that is obviously the use of the point-to-plane error metric which is the main difference to the

other algorithms all failing with the above specified accuracy demands. Interestingly, Wilm fails with rotatory registration for the torso phantom. Possibly the normal parameter of the point-to-plane error metric has disadvantages in this case. Similar happens for the AlignRT implementation, but for the bowl phantom.

5. Conclusion

In the paper, different unconstrained ICP algorithms have been compared for real (noisy) data produced by an optical sensor as part of a Tomotherapy HD system. Registration has to deal with mainly two difficulties: the deficiencies of the sensor (noise) and the ambiguities resulting from the shape of the measured object. Reference [3] found accuracies better than 1 mm and 0.5° for the used mannequin torso phantom with the proprietary registration software of the AlignRT system. We could show that such accuracies are only possible for well curved surfaces whereas gross errors may occur for registration of other not uniquely shaped surfaces and are not much effected by the chosen ICP registration algorithm.

The results show that obviously standard ICP algorithms only considering point cloud or surface data are too unreliable to serve as single verification tool of correct patient settlement. Of course, large correction values calculated by ICP registration give a clear hint that positioning is incorrect whereas the opposite case does not hold: as small value is no guarantee for correct alignment. Depending on the curvature of the actually captured surface parts, small ICP registration correction values are estimated even with wrong positioning because the ICP algorithm sticks in local minima. The registration information in parallel to the main orientation of the surface is only helpful in the case of unique surface structure. A safe registration useful for setup correction mostly yields perpendicular to the main orientation of the surface. Therefore, the result of ICP registration can only support the expertise of the clinical personnel as an additional tool for the positioning of the patient with regard to the treatment machine.

To improve the probability of reaching a correct deviation minimum without fiducial markers other variants of ICP algorithm including additional criteria such as colors, normals, and curvatures [25] may be applied. The hardware of the optical sensor supports this because an additional camera for capturing texture data is included in each measurement unit. But according to [3], although calibrated together with the stereo cameras, it can be only used for virtually projecting texture data on the captured surfaces, but not to support registration. Particularly, uncertainties of the registration in x-y direction could be reduced by this information.

Ongoing work is done on the estimation of confidence values of registration. Depending on curvatures characteristics of the treated regions an estimation of the reliability of a registration could be given. Also alternative registration approaches to surface registration, such as probabilistic methods [18], seem promising to improve the results and worthy of further investigation.

Competing Interests

The authors declare no conflict of interests.

Authors' Contributions

Gerald Krell, Nazila Nezhad, and Mathias Walke carried out the experiments, measured and collected the data and analyzed the results, and wrote the manuscript. Ayoub Al-Hamadi and Günther Gademann contributed to the experiments design and to the interpretation of results.

References

[1] J. Van Dyk, T. Kron, G. Bauman, and J. J. Battista, "Tomotherapy: a 'revolution' in radiation therapy," *Physics in Canada*, vol. 58, no. 2, pp. 79–86, 2002.

[2] H. M. Kjer and J. Wilm, *Evaluation of surface registration algorithms for PET motion correction [Ph.D. thesis]*, Technical University of Denmark, DTU, DK-2800 Kgs, Lyngby, Denmark, 2010.

[3] C. Bert, K. G. Metheany, K. Doppke, and G. T. Y. Chen, "A phantom evaluation of a stereo-vision surface imaging system for radiotherapy patient setup," *Medical Physics*, vol. 32, no. 9, pp. 2753–2762, 2005.

[4] P. J. Schöffel, W. Harms, G. Sroka-Perez, W. Schlegel, and C. P. Karger, "Accuracy of a commercial optical 3D surface imaging system for realignment of patients for radiotherapy of the thorax," *Physics in Medicine and Biology*, vol. 52, no. 13, pp. 3949–3963, 2007.

[5] G. Godin, M. Rioux, and R. Baribeau, "Three-dimensional registration using range and intensity information," in *Proceedings of the Processing and Analysis of 3D Data*, vol. 2350 of *Proceedings of SPIE*, pp. 279–290, Boston, Mass, USA, October 1994.

[6] M. Wiencierz, K. Kruppa, and L. Lüdemann, "Clinical validation of two surface imaging systems for patient positioning in percutaneous radiotherapy," https://arxiv.org/abs/1602.03749.

[7] T. Alderliesten, J.-J. Sonke, A. Betgen et al., "Accuracy evaluation of a 3-dimensional surface imaging system for guidance in deep-inspiration breath-hold radiation therapy," *International Journal of Radiation Oncology, Biology, Physics*, vol. 85, no. 2, pp. 536–542, 2013.

[8] L. I. Cerviño, T. Pawlicki, J. D. Lawson, and S. B. Jiang, "Frameless and mask-less cranial stereotactic radiosurgery: a feasibility study," *Physics in Medicine and Biology*, vol. 55, no. 7, pp. 1863–1873, 2010.

[9] A. P. Shah, T. Dvorak, M. S. Curry, D. J. Buchholz, and S. L. Meeks, "Clinical evaluation of interfractional variations for whole breast radiotherapy using 3-dimensional surface," *Practical Radiation Oncology*, vol. 3, no. 1, pp. 16–25, 2013.

[10] A. Fassi, J. Schaerer, M. Riboldi, D. Sarrut, and G. Baroni, "An image-based method to synchronize cone-beam CT and optical surface tracking," *Journal of Applied Clinical Medical Physics*, vol. 16, no. 2, 2015.

[11] D. B. Wiant, S. Wentworth, J. M. Maurer, C. L. Vanderstraeten, J. A. Terrell, and B. J. Sintay, "Surface imaging-based analysis of intrafraction motion for breast radiotherapy patients," *Journal of Applied Clinical Medical Physics*, vol. 15, no. 6, article 4957, 2014.

[12] T. Luhmann, "Close range photogrammetry for industrial applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 558–569, 2010.

[13] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV

cameras and lenses," in *Radiometry*, L. B. Wolff, S. A. Shafer, and G. Healey, Eds., pp. 221–244, Jones and Bartlett, Burlington, Mass, USA, 1992, http://dl.acm.org/citation.cfm?id=136913.136938.

[14] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proceedings of the 3rd International Conference on 3-D Digital Imaging and Modeling (3DIM '01)*, pp. 145–152, Quebec City, Canada, June 2001.

[15] D. W. Eggert, A. Lorusso, and R. B. Fisher, "Estimating 3-D rigid body transformations: a comparison of four major algorithms," *Machine Vision and Applications*, vol. 9, no. 5-6, pp. 272–290, 1997.

[16] K.-L. Low, *Linear Least-Squares Optimization for Point-to-Plane ICP Surface Registration*, Chapel Hill University of North Carolina, Chapel Hill, NC, USA, 2004.

[17] N. J. Mitra, N. Gelfand, H. Pottmann, and L. Guibas, "Registration of point cloud data from a geometric optimization perspective," in *Proceedings of the 2nd Symposium on Geometry Processing (SGP '04)*, pp. 22–31, Nice, France, July 2004.

[18] A. Myronenko and X. Song, "Point set registration: coherent point drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.

[19] Iterative Closest Point—File Exchange—MATLAB Central, http://www.mathworks.com/matlabcentral/fileexchange/27804-iterative-closest-point.

[20] Finite Iterative Closest Point—File Exchange—MATLAB Central, http://www.mathworks.com/matlabcentral/fileexchange/24301-finite-iterative-closest-point.

[21] The Simple GUI program for point clouds Registration—File Exchange—MATLAB Central, http://www.mathworks.com/matlabcentral/fileexchange/35019-the-simple-gui-program-for-point-clouds-registration.

[22] P. Bergström and O. Edlund, "Robust registration of point sets using iteratively reweighted least squares," *Computational Optimization and Applications*, vol. 58, no. 3, pp. 543–561, 2014.

[23] Iterative Closest Point Method—File Exchange—MATLAB Central, http://www.mathworks.com/matlabcentral/fileexchange/12627-iterative-closest-point-method.

[24] S. von Enzberg, E. Lilienblum, and B. Michaelis, "A physical simulation approach for active photogrammetric 3D measurement systems," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference (I2MTC '11)*, pp. 1–5, May 2011.

[25] D. Münch, B. Combés, and S. Prima, "A modified ICP algorithm for normal-guided surface registration," in *Proceedings of the Progress in Biomedical Optics and Imaging*, vol. 7623 of *Proceedings of SPIE*, San Diego, Calif, USA, February 2010.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 6740956, 12 pages
http://dx.doi.org/10.1155/2016/6740956

Research Article

A Fusion-Based Approach for Breast Ultrasound Image Classification Using Multiple-ROI Texture and Morphological Analyses

Mohammad I. Daoud,[1] Tariq M. Bdair,[1] Mahasen Al-Najar,[2] and Rami Alazrai[1]

[1]*Department of Computer Engineering, German Jordanian University, Amman, Jordan*
[2]*Jordan University Hospital, The University of Jordan, Amman, Jordan*

Correspondence should be addressed to Mohammad I. Daoud; mohammad.aldaoud@gju.edu.jo

Received 5 August 2016; Revised 31 October 2016; Accepted 15 November 2016

Academic Editor: Kenji Suzuki

Ultrasound imaging is commonly used for breast cancer diagnosis, but accurate interpretation of breast ultrasound (BUS) images is often challenging and operator-dependent. Computer-aided diagnosis (CAD) systems can be employed to provide the radiologists with a second opinion to improve the diagnosis accuracy. In this study, a new CAD system is developed to enable accurate BUS image classification. In particular, an improved texture analysis is introduced, in which the tumor is divided into a set of nonoverlapping regions of interest (ROIs). Each ROI is analyzed using gray-level cooccurrence matrix features and a support vector machine classifier to estimate its tumor class indicator. The tumor class indicators of all ROIs are combined using a voting mechanism to estimate the tumor class. In addition, morphological analysis is employed to classify the tumor. A probabilistic approach is used to fuse the classification results of the multiple-ROI texture analysis and morphological analysis. The proposed approach is applied to classify 110 BUS images that include 64 benign and 46 malignant tumors. The accuracy, specificity, and sensitivity obtained using the proposed approach are 98.2%, 98.4%, and 97.8%, respectively. These results demonstrate that the proposed approach can effectively be used to differentiate benign and malignant tumors.

1. Introduction

Breast cancer is the most common cancer in women worldwide and one of the major causes of death in females across the globe [1]. The statistics of the World Health Organization (WHO) indicate that, in 2012, 1.67 million new cases were diagnosed with breast cancer and around 522,000 women died of this disease [1]. Early diagnosis of breast cancer is crucial for the successful treatment of the disease and improving the survival rates of the patients [2].

Ultrasound imaging is one of the most widely used imaging modalities for breast cancer diagnosis since it offers the advantages of low-cost, portability, patient comfort, and diagnosis accuracy [3, 4]. However, the interpretation of breast ultrasound (BUS) images is operator-dependent and varies based on the experience and skill of the radiologist [5]. To overcome this limitation, computer-aided diagnosis (CAD) systems have been introduced to analyze BUS images and provide the radiologist with a second opinion to improve the diagnosis accuracy and reduce the effect of operator dependency [5, 6].

Many studies, such as [7–15], have employed BUS image analysis for classifying breast tumors. In particular, morphological features [13, 16, 17] and texture features [8, 12] are demonstrated to be useful for differentiating benign and malignant tumors. Moreover, combining both feature groups has been suggested to improve the tumor classification accuracy [13, 18]. Morphological features quantify the geometrical characteristics of the tumor, such as area, shape, orientation, regularity, and margins [6, 19]. Therefore, morphological features are mainly affected by the accuracy of the tumor outline. Commonly used morphological feature descriptors

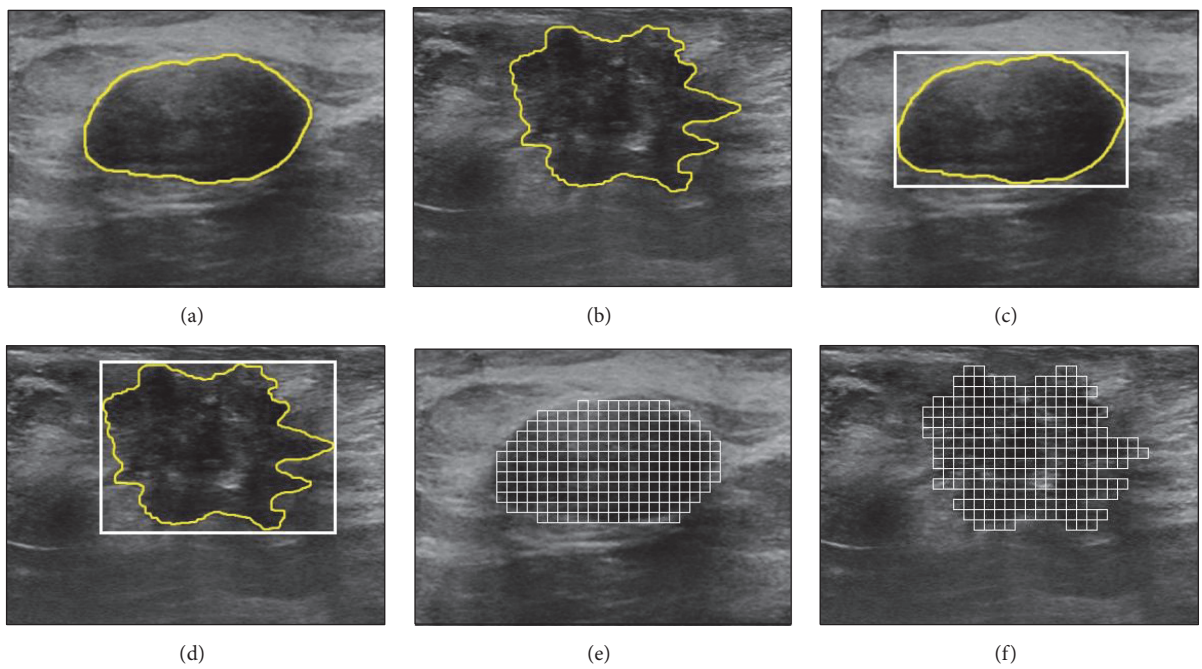


Figure 1: (a)-(b) BUS images of (a) benign and (b) malignant tumors with manually drawn outlines (yellow line). (c)-(d) A single ROI is drawn around each tumor in (a) and (b), such that the ROI corresponds to the minimum bounding rectangle that contains the tumor. Such ROI is often used in conventional texture analysis. (e)-(f) Each tumor in (a) and (b) is divided into a set of nonoverlapping ROIs. These multiple ROIs are used in the proposed approach to extract the texture features.

include the aspect ratio [13, 17], the best-fit ellipse of the tumor, the normalized radial length (NRL) [18, 20], and the undulation characteristics [21].

Texture features quantify the pixel gray-level statistics in terms of intensity and spatial distribution [6]. Generally, the texture patterns of benign tumors are different from those of malignant tumors [10]. Therefore, several texture descriptors have been employed for classifying BUS images [22–26]. Among these descriptors, the gray-level cooccurrence matrix (GLCM) [27] is one of the most widely used texture analysis techniques for BUS image classification [12]. Conventional texture analysis often uses a single region of interest (ROI) to extract global texture features that quantify the texture characteristics of the entire tumor. One of the most common ROI selection procedures is to find the minimum bounding rectangle that encloses the tumor [9, 12, 22]. Another ROI selection approach is to find the maximum rectangle that fits inside the tumor [28]. Such ROIs can be drawn manually by a radiologist or detected automatically using a computer algorithm.

In many BUS images, the local texture patterns within the tumor vary from one region to another. Hence, the use of a single ROI, which enables the extraction of global texture features that quantify the entire tumor, might not support effective quantification of the local texture variations within the tumor. Moreover, the mismatch between the predefined structure of the ROI and the actual shape of the tumor might reduce the tumor classification accuracy. For example, consider the benign and malignant tumors shown in Figures 1(a) and 1(b), respectively. The texture patterns inside each tumor demonstrate local variations. For both tumors, the ROIs corresponding to the minimum bounding rectangle that encloses the tumor are presented in Figures 1(c) and 1(d). Both ROIs might not provide efficient extraction of texture features that can effectively quantify the local texture variations within the tumor. In addition, the ROI of each tumor extends beyond the tumor boundary, and hence the texture features extracted from such ROI are expected to quantify both the tumor and the surrounding healthy tissue. These limitations might lead to imprecise texture analysis of the tumor, which in turn can reduce the tumor classification accuracy.

To improve the tumor classification capability of ultrasound texture analysis, this study investigates the use of multiple ROIs to analyze the local pixel gray-level statistics inside the tumor. In particular, the tumor is divided into a set of nonoverlapping ROIs as illustrated in Figures 1(e) and 1(f). Each ROI is analyzed individually to extract local texture features. The texture features employed in this study are computed using the GLCM matrix. A local tumor class indicator is estimated for each individual ROI by classifying the texture features of that ROI using a well-trained classifier. The class of the tumor can be determine based on the multiple-ROI texture analysis by employing a majority voting mechanism to integrate the local tumor class indicators of all ROIs inside the tumor. The proposed multiple-ROI texture analysis approach enables effective quantification of the local texture patterns inside the tumor without incorporating

texture patterns of the healthy tissue that surrounds the tumor.

One challenge of applying the proposed multiple-ROI texture analysis approach is to enable effective combination between the local texture features, which are extracted for each one of the multiple ROIs inside the tumor, with the morphological features that are computed for the entire tumor. Therefore, a novel probabilistic approach is proposed to fuse the tumor classification indicators obtained using the multiple-ROI texture analysis with the tumor classification indicator computed using morphological analysis of the entire tumor. The morphological analysis employed in this paper is based on set of morphological features introduced in previous studies [13, 17, 18, 20, 21, 29] to quantify the shape and contour of the tumor.

To evaluate the performance of the proposed BUS image classification approach, both the multiple-ROI texture analysis and the fusion-based combination between the multiple-ROI texture analysis and morphological analysis are employed to classify a BUS image database that includes 64 benign tumors and 46 malignant tumors. These BUS images were acquired during ultrasound breast cancer screening procedures. The tumor classifications results of the proposed approach are compared with conventional texture (single ROI), morphological, and combined texture and morphological analyses.

The remainder of the paper is organized as follows. The data acquisition of the BUS image database is summarized in Section 2. Moreover, Section 2 describes the conventional texture and morphological analysis of BUS images, the proposed tumor classification approach, and the performance metrics employed to compare the conventional and proposed BUS image classification approaches. The experimental results and discussion are provided in Section 3. Finally, the conclusion is presented in Section 4.

2. Materials and Methods

2.1. Data Acquisition. The collected image database consists of 110 BUS images of pathologically proven benign and malignant tumors (64 benign tumors and 46 malignant tumors). Detailed description of the types of benign and malignant tumors involved in this study is provided in Table 1. Each BUS image was acquired from one patient (i.e., the number of patients which participated in the study is 110). All participated patients were females. Moreover, each image included exactly one breast tumor. The age of the patients ranged from 25 to 77 years. The mean and standard deviation of the maximum diameters of the tumors are 14.7 mm and 6.0 mm, respectively. The BUS images were acquired during routine ultrasound breast cancer screening procedures at the Jordan University Hospital, Amman, Jordan, during the period between May 2012 and February 2016. Ultrasound imaging was performed using an Acuson S2000 ultrasound system (Siemens AG, Munich, Germany) and a 14L5 linear transducer with frequency bandwidth from 5 to 14 MHz. During imaging, the radiologist was free to adjust the configurations of the imaging system, including the focal length, depth, and gain to obtain the best view. For each BUS image, the tumor was manually outlined by a radiologist with more than 13 years of experience. The tumor outlines were also verified by another independent experienced radiologist. All images were resampled to the same resolution of 0.1 mm × 0.1 mm per pixel. The study protocol was approved by the ethics committee at the Jordan University Hospital. Moreover, informed consent to the protocol was obtained from each patient.

Table 1: Description of the benign and malignant breast tumors involved in the study.

Tumor class	Description	Number of patients
Benign	Fibroadenoma	35
	Complex fibroadenoma	1
	Fibrocystic change	15
	Chronic inflammation	1
	Lymphocytic lobulitis	1
	Fibrosis	3
	Sclerosing adenosis	1
	Atypical ductal hyperplasia	1
	Atypical lobular hyperplasia	1
	Adenosis	2
	Chronic mastitis	1
	Tubular adenoma	1
	Fat necrosis	1
Malignant	Invasive ductal carcinoma	41
	Ductal carcinoma in situ	4
	Invasive lobular carcinoma	1

2.2. Quantitative Features. Both texture and morphological features are used to classify the benign and malignant breast tumors. The following two sections describe both feature groups.

2.2.1. Texture Features. The texture features employed in this study were computed using the GLCM matrix [27], which measures the correlations between adjacent pixels within a ROI. The computation of the GLCM matrix was performed using four distances (d = 1, 2, 3, and 4 pixels) and four different orientations ($\theta = 0^\circ, 45^\circ, 90^\circ$, and 135°). Therefore, sixteen GLCM matrices were computed for each ROI. Each GLCM matrix was analyzed, as described in [12], to extract twenty texture features (TF1–TF20) that are commonly used for ultrasound texture analysis [12, 32]. These texture features are provided in Table 2. Thus, a total of 320 texture features were extracted from each ROI.

2.2.2. Morphological Features. In this study, eighteen morphological features are extracted from each tumor. Among these features, ten features can be extracted directly from the tumor (MF1–MF10). Six morphological features are extracted from the best-fit ellipse that approximates the size and position of the tumor (MF11–MF16). The last two morphological features are the entropy (MF17) and variance (MF18) of the

Table 2: The morphological and texture features employed for tumor classification.

Category	Feature	Code	Description
Texture	Autocorrelation [30]	TF1	Twenty texture features (TF1–TF20) are extracted from GLCM matrices computed using four distances ($d = 1, 2, 3, 4$ pixels) and four orientations ($\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$)
	Contrast [12]	TF2	
	Correlation [30]	TF3	
	Cluster prominence [30]	TF4	
	Cluster shade [30]	TF5	
	Dissimilarity [30]	TF6	
	Energy [30]	TF7	
	Entropy [30]	TF8	
	Homogeneity [30]	TF9	
	Maximum probability [30]	TF10	
	Sum of squares [27]	TF11	
	Sum average [27]	TF12	
	Sum entropy [27]	TF13	
	Sum variance [27]	TF14	
	Difference variance [27]	TF15	
	Difference entropy [27]	TF16	
	Information measure of correlation I [27]	TF17	
	Information measure of correlation II [27]	TF18	
	Inverse difference normalized [31]	TF19	
	Inverse difference moment normalized [31]	TF20	
Morphological	Tumor area [20]	MF1	Ten morphological features (MF1–MF10) are extracted directly from the tumor
	Perimeter [20]	MF2	
	Form factor [13, 17]	MF3	
	Roundness [13, 17]	MF4	
	Aspect ratio [13, 17]	MF5	
	Convexity [13, 17]	MF6	
	Solidity [13, 17]	MF7	
	Extent [13, 17]	MF8	
	Undulation characteristics [21]	MF9	
	Compactness [20, 29]	MF10	
Morphological	Length of the ellipse major axis [20]	MF11	Six morphological features (MF11–MF16) are extracted from the best-fit ellipse that approximates the size and position of the tumor
	Length of the ellipse minor axis [20]	MF12	
	Ratio between the ellipse major and minor axes [20]	MF13	
	Ratio of the ellipse perimeter and the tumor perimeter [20]	MF14	
	Overlap between the ellipse and the tumor [20]	MF15	
	Angle of the ellipse major axis [20]	MF16	
Morphological	NRL entropy [18, 20]	MF17	Two morphological features (MF17-MF18) are extracted from the NRL of the tumor
	NRL variance [18, 20]	MF18	

normalized radial length (NRL) of the tumor [18, 20]. The NRL is defined as the distance between the tumor center and the pixels located on the tumor boundary normalized to the maximum radial length of the tumor [18]. The eighteen morphological features are summarized in Table 2.

2.3. Conventional Tumor Classification. The 110 BUS images are analyzed using conventional tumor classification analysis, as illustrated in Figure 2. In particular, the GLCM texture features, described in Section 2.2.1, are extracted from a single ROI. As mentioned in the Introduction, this ROI corresponds to the minimum bounding rectangle that encloses the tumor. The morphological features, summarized in Section 2.2.2, are extracted from the outlined tumor.

Feature selection, which eliminates the irrelevant and redundant features, is applied to determine the best subsets

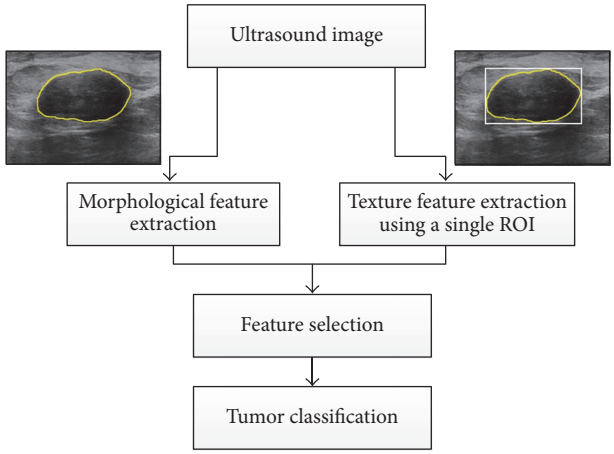


Figure 2: Overview of conventional tumor classification in which texture features are extracted from a single ROI that encloses the tumor and morphological features are computed based on the tumor outline. Both groups of features are processed using feature selection and classification to differentiate benign and malignant tumors.

of texture, morphological, and combined texture and morphological features that reduce the misclassification error between malignant and benign tumors. In fact, exhaustive search for the optimal feature combination requires extensive computational resources and long processing times, particularly when the number of features is large. For example, the total number of all potential combinations of n features into m subsets is equal to $(1/m!)\sum_{i=0}^{m}(-1)^{m-i}\binom{m}{i}i^n$ [33]. Therefore, a two-phase heuristic approach, which is based on the feature selection procedures described in [12, 34], is employed to carry out feature selection. In the first phase, the features are ranked according to the minimal-redundancy-maximal-relevance (mRMR) criterion [34], which is based on mutual information. The top l-ranked features are incrementally grouped and their classification performance is evaluated, for all $l = \{1, 2, \ldots, L\}$, where L is the total number of features. The smallest feature group that can achieve the minimum classification error is taken as the candidate feature subset. In the second phase, the backward selection algorithm is applied to the candidate feature subset. In this algorithm, features are sequentially eliminated until the removal of further features leads to degrading the classification accuracy. This two-phase algorithm enables the selection of a compact feature subset that can achieve effective tumor classification.

The selected features are classified using a binary SVM classifier [35] that is implemented using the LIBSVM library [36]. In binary SVM, the input features are mapped into a high dimensional feature space by applying a kernel function. This mapping enables the computation of a nonlinear decision function that can separate the feature space into two regions, one for each class. Specifically, given a training set $T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_k \in R^N$ represents the kth feature vector and $y_k \in \{-1, +1\}$ is the corresponding tumor class. The goal of SVM is to determine a decision boundary in the form of hyperplane that can separate the feature space into two regions through maximizing the margin between the samples of different classes. The resultant decision function is defined as follows:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{k=1}^{n} y_k \alpha_k \varphi(\mathbf{x}_k, \mathbf{x}) + b\right), \quad (1)$$

where $\mathbf{x} \in R^N$ is a new feature vector to be classified into benign or malignant, $\varphi(\mathbf{x}_k, \mathbf{x})$ is a kernel function that maps the input vectors into high dimensional space, α_k is the kth Lagrange multiplier, and b is the bias term of the decision hyperplane. Several kernel functions can be used with SVM. However, the Gaussian radial basis function (RBF) is by far the most commonly used kernel function for classification tasks [37]. In this work, the RBF kernel is employed. The RBF kernel function can be defined as follows:

$$\varphi(\mathbf{x}_k, \mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}_k - \mathbf{x}\|^2}{2\sigma^2}\right), \quad (2)$$

where $\sigma > 0$ is the RBF kernel parameter.

The performance of the SVM classifier with RBF kernel depends on two parameters: σ, the RBF kernel parameter, and $C > 0$, the regularization parameter. The tuning of the two parameters is carried out using a grid-based search of the two-dimensional parameter space $1 < \sigma < 100$ and $1 < C < 100$. The search is performed with a step length of 1. The best SVM model is selected such that its parameters maximize the average tumor classification accuracy.

The performance evaluation of the conventional tumor classification is performed using the single ROI GLCM texture features, the morphological features, and the combined single ROI texture features and morphological features. Similar to the work of Wu et al. [13], the evaluation is carried out using a fivefold cross-validation procedure. In this procedure, 80% of the tumors are selected for training and the remaining 20% is used for testing. This process is repeated five times so that each of the 110 BUS images is included once in the testing.

2.4. The Proposed Tumor Classification Approach. The architecture of the proposed tumor classification approach is illustrated in Figure 3. In this architecture, the multiple-ROI texture analysis is carried out by dividing the tumor into small, nonoverlapping ROIs and extracting local texture features from each individual ROI. Moreover, the tumor is analyzed to extract morphological features. To combine the local texture features of the individual ROIs and the global morphological features, two independent posterior tumor class likelihoods are obtained separately from the multiple-ROI texture analysis and the morphological analysis. Moreover, decision fusion is applied to fuse both tumor class likelihoods and determine the class of the tumor.

To perform the multiple-ROI texture analysis, the tumor is divided into a set of uniform, nonoverlapping ROIs, as shown in Figures 1(e) and 1(f). The size of the ROIs is estimated by considering three factors: preserving the capability of differentiating various texture patterns, reducing the possibility of including different local textures within the

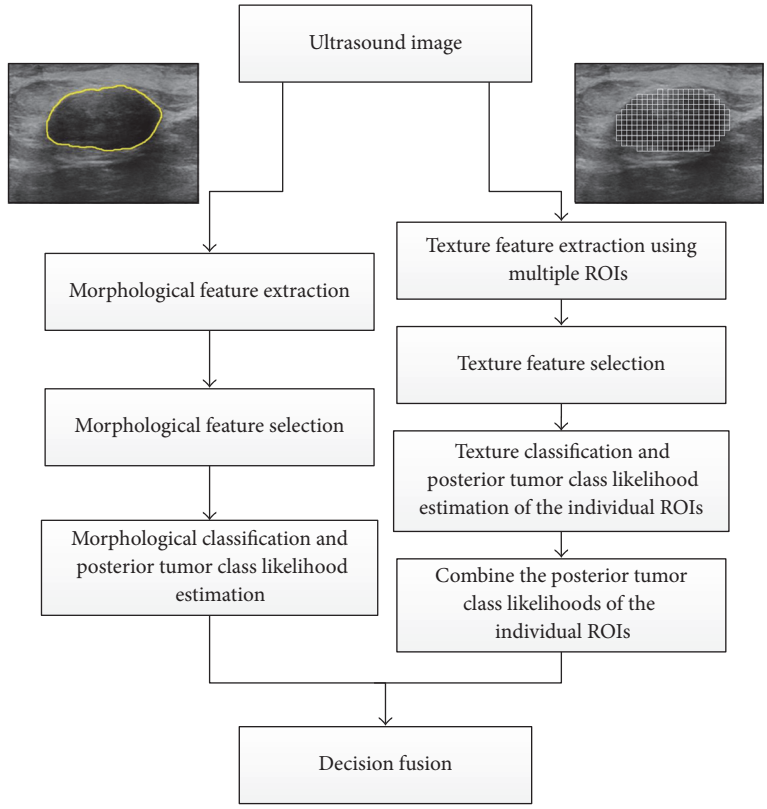


Figure 3: The architecture of the proposed tumor classification approach. Texture analysis is performed by dividing the tumor into a group of nonoverlapping ROIs and extracting texture features from each ROI. A selected set of texture features are used to classify each individual ROI and compute its posterior tumor class likelihood. The posterior tumor class likelihoods of the individual ROIs are combined. Morphological analysis is performed by extracting morphological features from the outlined tumor and employing a selected set of the features to predict the posterior tumor class likelihood. Decision fusion is then used to combine the posterior tumor class likelihoods obtained using the texture and morphological analyses and determine the tumor class.

same ROI, and ensuring that the entire tumor is adequately covered by the ROIs. The study by Valckx and Thijssen [38] suggested that the use of very small ROIs might degrade the capability of differentiating various texture patterns. On the other hand, the use of large ROIs increases the possibility of including different local texture patterns within a single ROI. Moreover, the use of large ROIs might lead to big gaps: that is, areas that are not covered by the ROIs, at the tumor boundary. For example, consider Figures 4(a), 4(c), and 4(e) that show the benign tumor in Figure 1(a) divided into uniform ROIs of size $0.5 \times 0.5\,\text{mm}^2$, $1 \times 1\,\text{mm}^2$, and $2 \times 2\,\text{mm}^2$, respectively. Moreover, consider Figures 4(b), 4(d), and 4(f) that show the malignant tumor in Figure 1(b) divided into ROIs of sizes $0.5 \times 0.5\,\text{mm}^2$, $1 \times 1\,\text{mm}^2$, and $2 \times 2\,\text{mm}^2$, respectively. The use of the $0.5 \times 0.5\,\text{mm}^2$ ROIs minimizes the possibility of including different local textures within a single ROI and reduces the gaps at the tumor boundary. However, the small size of the ROIs, which corresponds to 5×5 pixels, might limit the ability of the texture analysis to differentiate various texture patters. On the other hand, the use of the $2 \times 2\,\text{mm}^2$ ROIs, which correspond to 20×20 pixels, enables better texture classification but increases the possibilities of including different local textures within the same ROI and producing large gaps at the tumor boundary. The $1 \times 1\,\text{mm}^2$ ROIs, which correspond to 10×10 pixels, provide a reasonable balance between the need to use ROIs of reasonable size to enable effective texture analysis and the requirements of reducing the possibility of crossing different local textures within a single ROI and achieving adequate coverage of the entire tumor. Hence, the size of the ROIs employed in this study is set to $1 \times 1\,\text{mm}^2$.

Each ROI is processed individually to extract the GLCM texture features described in Section 2.2.1. The two-phase feature selection algorithm described in Section 2.3 is employed to determine the subset of texture features that enables the best tumor classification accuracy based on the multiple-ROI texture analysis. A binary SVM classifier with RBF kernel is used to classify each ROI as benign or malignant using the selected subset of texture features. The tuning of the SVM parameters is achieved using the grid-based search described in Section 2.3. The posterior tumor class likelihood of each ROI is estimated from the SVM output using Platt's approach [39]. Then, a majority voting mechanism is used to determine the class of the tumor based on the classification indictors of the individual ROIs. In particular, if more than 50% of the ROIs in the tumor are classified as malignant, then the tumor

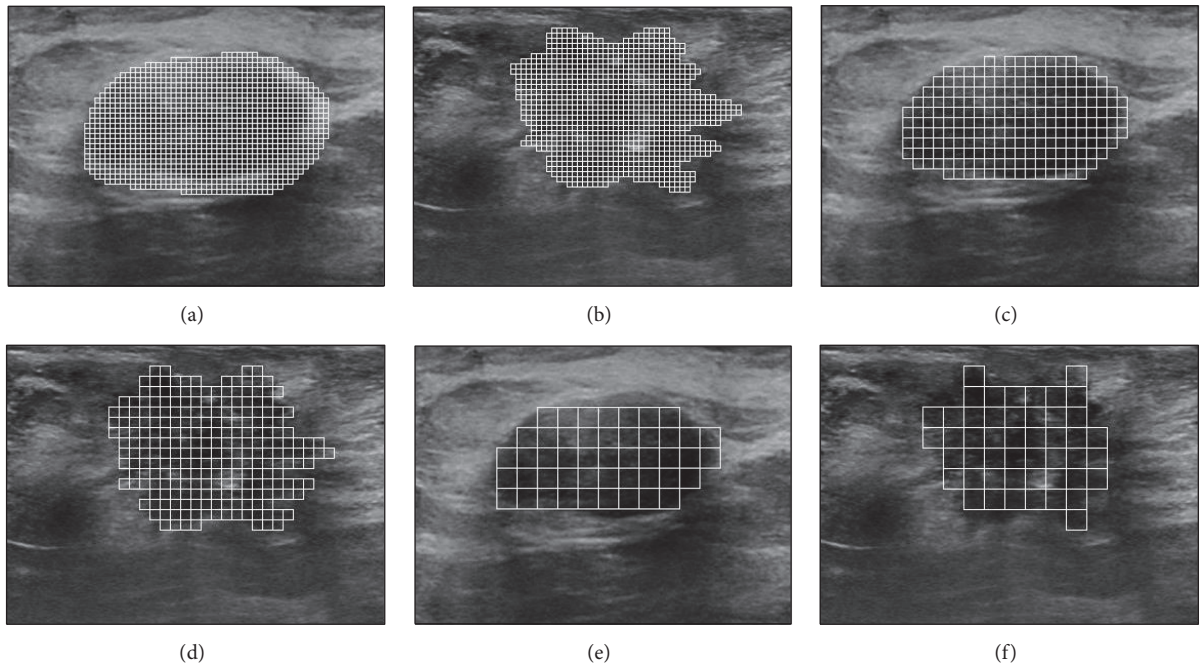


Figure 4: The benign and malignant tumors shown in Figures 1(a) and 1(b), respectively, are divided into a set of nonoverlapping ROIs with a size of (a)-(b) $0.5 \times 0.5\,\text{mm}^2$, (c)-(d) $1 \times 1\,\text{mm}^2$, and (e)-(f) $2 \times 2\,\text{mm}^2$.

is considered malignant. Otherwise, the tumor is considered benign. The computation of the posterior likelihood of the tumor is performed by averaging the posterior tumor class likelihoods of the ROIs that agree with the class of tumor estimated using the multiple-ROI texture analysis.

To perform the morphological analysis, the extraction and selection of the morphological features as well as the tuning of the SVM classifier match those of the conventional morphological-based classification that was described in Section 2.3. Moreover, the tuned SVM is used to classify the tumor based on the selected morphological features and Platt's approach is applied to compute the posterior tumor class likelihood of the entire tumor.

For a given BUS image, the posterior tumor class likelihood obtained using the multiple-ROI texture analysis is mutually independent from the posterior tumor class likelihood estimated using the morphological analysis. Therefore, the fusion of the tumor class decisions obtained using these two independent analyses can be performed using a Gaussian Naive-Bayes approach [40].

To apply the Gaussian Naive-Bayes approach, consider a vector of continuous decisions $\mathbf{D} = [d_1, \ldots, d_L]^T$ obtained from L different classifiers for a specific BUS image. The probability that the BUS image belongs to class y given decisions of the L different classifiers can be written as

$$P(y \mid d_1, \ldots, d_L) = \frac{P(y)\, P(d_1, \ldots, d_L \mid y)}{P(d_1, \ldots, d_L)}, \tag{3}$$

where for binary classification, which is considered in this study, $y \in \{-1, 1\}$ and $L = 2$. Using the mutual independence assumption between the two classifiers, (3) can be rewritten as

$$P(y \mid d_1, \ldots, d_L) = \frac{P(y) \prod_{i=1}^{L} P(d_i \mid y)}{P(d_1, \ldots, d_L)}. \tag{4}$$

The term $P(d_1, \ldots, d_L)$ is a normalization factor. Therefore, a BUS image can be classified based on the combined decisions from the $L = 2$ classifiers using the following decision rule:

$$\widehat{y} = \arg\max_{y} \left(P(y) \prod_{i=1}^{L} P(d_i \mid y) \right), \tag{5}$$

where $P(d_i \mid y)$ is assumed to be a multivariant normal distribution with mean vector μ_i and covariance matrix $C_i \in R^{L \times L}$. The class prior probability $P(y)$ and the parameters (μ_i, C_i) are estimated using maximum likelihood [41].

The performance evaluation of the proposed tumor classification approach is carried out using two different configurations. In the first configuration, the tumor is classified using the multiple-ROI texture analysis only. In the second configuration, tumor classification is carried out by fusing the posterior tumor class likelihoods of the multiple-ROI texture analysis and the morphological analysis. In both configurations, the fivefold cross-validation procedure described in Section 2.3 is employed. It is worth noting that the selection of the ROIs during the fivefold SVM training and testing of the multiple-ROI texture analysis was tumor-specific. In other words, in each fold of the cross-validation procedure, the training was performed using ROIs that belong to 80% of the tumors, while the testing was carried out with the ROIs of the remaining 20% of the tumors.

Table 3: Classification results of the 110 BUS images obtained using the proposed approach.

BUS image classification	Multiple-ROI texture analysis		Fusion of the multiple-ROI texture analysis and the morphological analysis	
	Benign*	Malignant*	Benign*	Malignant*
Benign	60 TN	1 FN	63 TN	1 FN
Malignant	4 FP	45 TP	1 FP	45 TP
Total	64	46	64	46

*Histological finding.

2.5. Performance Evaluation. Six objective metrics, namely, the accuracy, specificity, sensitivity, negative predictive value (NPV), positive predictive value (PPV), and Matthew's correlation coefficient (MCC) [6], are used to evaluate the performance of the conventional tumor classification as well as the proposed tumor classification. These metrics are defined as follows:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},\\
\text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}},\\
\text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}},\\
\text{PPV} &= \frac{\text{TP}}{\text{TP} + \text{FP}},\\
\text{NPV} &= \frac{\text{TN}}{\text{TN} + \text{FN}},\\
\text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}},
\end{aligned} \tag{6}$$

where TP is the number of true positive cases, TN is the number of true negative cases, FP is the number of false positive cases, and FN is the number of false negative cases.

The relationships between specificity and sensitivity, achieved using the conventional and proposed classification approaches, are analyzed by drawing the receiver operator characteristic (ROC) curves. Moreover, the area under the ROC curve (AUC), which quantifies the overall performance of a CAD system, is computed for each classification approach.

To confirm the effectiveness of the proposed fusion-based approach, paired t tests were carried out on average classification accuracies to compare the fused multiple-ROI texture and morphological analyses with the other four classification approaches.

The execution times of the conventional texture, morphological, and combined texture and morphological analyses are compared with the proposed multiple-ROI texture analysis and the fused multiple-ROI texture and morphological analyses. The compression was performed by implementing the five approaches using MATLAB (MathWorks Inc., Natick, Massachusetts, USA) and executing them on a computer workstation that has a 3.5 GHz processor and 16 GB of memory and runs Ubuntu Linux operating system. For each one of the five classification approaches, the total time required to extract the features and classify the BUS image was recorded for twenty trials.

3. Results and Discussion

The tuned values of the SVM parameters (σ, C) that are used to carry out tumor classification using the conventional texture features, morphological features, and combined texture and morphological features are equal to (3,56), (3,50), and (2,50), respectively. Moreover, the tuned values of (σ, C) that are employed to perform tumor classification using the proposed multiple-ROI texture analysis are equal to (4,55). To carry out the fusion-based tumor classification, both the multiple-ROI texture analysis and the morphological analysis are performed using their optimized SVM parameters (i.e., the parameters (4,55) are used for the multiple-ROI texture analysis and (3,50) are employed for the morphological analysis).

The features selected to perform the proposed multiple-ROI texture analysis are $\text{TF1}_{4,90^\circ}$, $\text{TF2}_{3,90^\circ}$, $\text{TF4}_{4,90^\circ}$, $\text{TF6}_{4,90^\circ}$, $\text{TF8}_{4,90^\circ}$, $\text{TF9}_{4,90^\circ}$, $\text{TF10}_{3,90^\circ}$, $\text{TF11}_{3,45^\circ}$, $\text{TF12}_{3,45^\circ}$, $\text{TF13}_{3,135^\circ}$, $\text{TF14}_{4,45^\circ}$, $\text{TF15}_{2,90^\circ}$, $\text{TF16}_{2,90^\circ}$, $\text{TF17}_{4,90^\circ}$, and $\text{TF18}_{4,90^\circ}$, where the first subscript is the distance, d, and the second is the orientation angle, θ. The proposed fusion-based tumor classification was performed using the aforementioned multiple-ROI texture features as well as the selected subset of morphological features. These morphological features are MF1, MF2, MF3, MF4, MF5, MF6, MF7, MF8, MF13, MF14, and MF18.

The results achieved by the proposed tumor classification approach using the multiple-ROI texture analysis as well as the fused multiple-ROI texture and morphological analyses are shown in Table 3 with respect to the pathological findings. Both configurations of the proposed approach achieved effective classification of benign and malignant breast tumors. However, the fusion of the multiple-ROI texture analysis and morphological analysis enabled higher classification performance than that obtained using the multiple-ROI texture analysis alone.

The six objective performance metrics obtained for the proposed classification approach and conventional classification approach are presented in Table 4. The conventional classification approach achieved better performance by combining the texture and morphological features than that

Table 4: Objective performance metrics obtained using the (a) conventional classification approach using texture features, (b) conventional classification approach using morphological features, (c) conventional classification approach using both texture and morphological features, (d) proposed classification approach using multiple-ROI texture analysis, and (e) proposed classification approach using the fused multiple-ROI texture analysis and morphological analysis.

	(a)	(b)	(c)	(d)	(e)
Accuracy	85.5%	87.3%	90.9%	95.5%	98.2%
Specificity	84.4%	89.1%	90.6%	93.8%	98.4%
Sensitivity	87.0%	84.8%	91.3%	97.8%	97.8%
PPV	80.0%	84.8%	87.5%	91.8%	97.8%
NPV	90.0%	89.1%	93.6%	98.4%	98.4%
MCC	70.7%	73.9%	81.5%	90.9%	96.3%

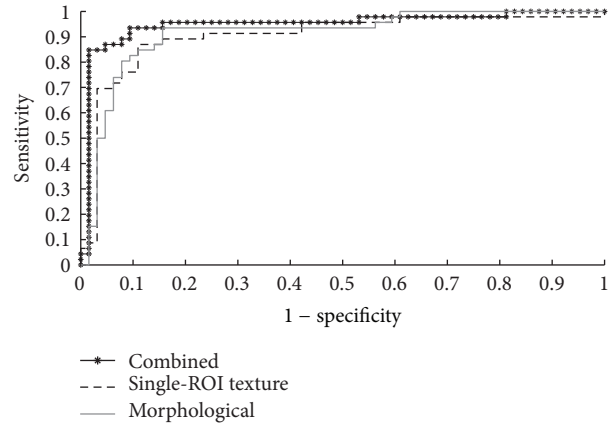


Figure 5: The ROC curves of the conventional classification approach using texture features, morphological features, and the combined texture and morphological features.

obtained by only using the texture features or the morphological features. This finding agrees with the results reported in previous studies [13, 14]. Moreover, the classification results demonstrate that the proposed approach using the multiple-ROI texture analysis outperforms the conventional classification using the texture, morphological, and combined texture and morphological features. In particular, the multiple-ROI texture analysis achieved classification accuracy of 95.5%, specificity of 93.8%, sensitivity of 97.8%, PPV of 91.8%, NPV of 98.4%, and MCC of 90.9%. The optimal classification performance was achieved by the proposed approach using the fused multiple-ROI texture analysis and morphological analysis. Specifically, the fusion of the multiple-ROI texture and morphological analyses enabled classification accuracy of 98.2%, specificity of 98.4%, sensitivity of 97.8%, PPV of 97.8%, NPV of 98.4%, and MCC of 96.3%.

The ROC curves of the conventional classification approach and the proposed classification approach are shown in Figures 5 and 6, respectively. The AUC values obtained for the conventional classification using the texture features, morphological features, and combined texture and morphological features are equal to 0.902, 0.912, and 0.948, respectively. The proposed classification approach achieved AUC values of 0.963 using the multiple-ROI texture analysis and 0.975 using the fused multiple-ROI texture and morphological analyses. These results confirm the superior performance of the proposed classification approach compared to conventional BUS image classification.

The p values obtained using the paired t tests to compare the proposed fused multiple-ROI texture and morphological analyses with the other four classification approaches at confidence level of 0.05 are shown in Table 5. The results reported in Table 5 demonstrate that the fusion-based approach outperforms significantly the conventional classification using the texture features, morphological features, and combined texture and morphological features as well as the multiple-ROI texture analysis.

According to these results, our proposed tumor classification approach achieved high sensitivity of 97.8% using both the multiple-ROI texture analysis and the fused multiple-ROI

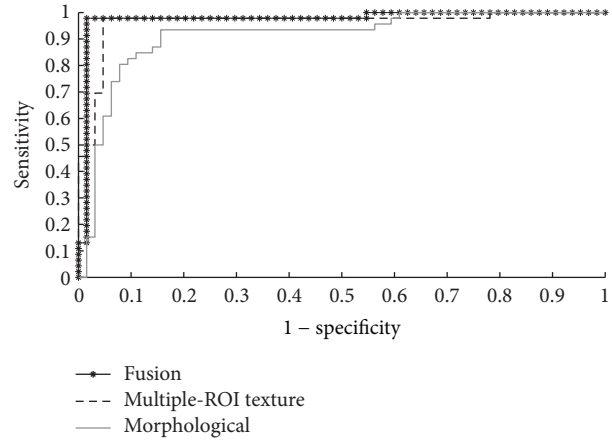


Figure 6: The ROC curves of the proposed classification approach using the multiple-ROI texture analysis, the morphological analysis, and the fused multiple-ROI texture analysis and morphological analysis.

Table 5: Comparisons of the p values computed using paired t-tests on average accuracies between the fused multiple-ROI texture and morphological analyses and the (a) conventional classification approach using texture features, (b) conventional classification approach using morphological features, (c) conventional classification approach using both texture and morphological features, and (d) multiple-ROI texture analysis.

	(a)	(b)	(c)	(d)
p value	0.007	0.011	0.041	0.046

texture and morphological analyses. Such finding suggests that the proposed approach enables high probability of diagnosing malignant tumors. Moreover, the near-perfect values of PPV and NPV obtained by fusing the multiple-ROI texture analysis and morphological analysis indicate that the number of unnecessary biopsies carried out for

benign tumors can be minimized. These results also suggest that the proposed approach has the potential to provide the radiologists with a second opinion that effectively reduces the rate of misdiagnosis.

The mean ± standard deviation execution times of the multiple-ROI texture analysis and the fused multiple-ROI texture and morphological analyses are 72.20 ± 2.14 s and 73.66 s ± 2.19 s, respectively. In comparison, the mean ± standard deviation execution times of the conventional texture, morphological, and combined texture and morphological analyses are 0.16 ± 0.03 s, 1.47 ± 0.18 s, and 1.63 ± 0.19 s, respectively. Although the multiple-ROI texture analysis and the fused multiple-ROI texture and morphological analyses are slower than the conventional classification analyses, both proposed classification approaches require around one minute to classify the BUS image. Such execution times do not limit the application of the proposed classification approaches in CAD systems that aim to provide an accurate second opinion to the radiologist.

The results reported in this study indicate that the proposed multiple-ROI texture analysis outperforms the conventional texture analysis in which texture features are extracted from a single ROI that includes the tumor. As mentioned in the Introduction, many breast tumors might have complicated texture patterns that vary from one region to another inside the tumor. Therefore, the multiple-ROI texture analysis enables effective quantification of the different local texture patterns inside the tumor. Another factor that might contribute to the improved performance of the multiple-ROI texture analysis is its ability to analyze the local texture patterns of the tumor without incorporating texture patterns of the surrounding healthy tissue.

The use of small ROIs for tissue characterization has been employed by other ultrasound-based methods. For example, in quantitative ultrasound imaging of cancer [42, 43], the raw ultrasound radio-frequency (RF) signals are divided into small ROIs, and each ROI is analyzed to extract spectral features for tissue characterization. Moreover, a recent study by Uniyal et al. [44] has compared the classification performance of a combination of ultrasound-based texture, spectral, and RF time series features that are extracted from the entire breast tumor with the performance obtained by dividing the tumor into $1 \times 1\,\text{mm}^2$ ROIs and extracting similar ultrasound-based features from each individual ROI. This study demonstrates that the classification performance obtained by classifying the individual $1 \times 1\,\text{mm}^2$ ROIs outperforms the classification results achieved by classifying the entire tumor. This finding agrees with our proposed multiple-ROI texture analysis approach.

The multiple-ROI texture analysis has been applied in the current study to improve the classification performance of GLCM texture features. Our future directions include extending the multiple-ROI texture analysis approach to incorporate other statistical texture methods that use a ROI to extract texture features. The proposed approach can also be extended by performing multiresolution texture feature extraction, in which ROIs of different sizes are employed to carry out the multiple-ROI texture analysis. Moreover, the probabilistic approach, which has been used in this study to fuse the multiple-ROI texture analysis with the morphological analysis, can be expanded to support the fusion of multiple classification results obtained using various texture and morphological methods with the goal of achieving higher accuracy, specificity, and sensitivity.

One important factor that affects the tumor classification performance is the ability to accurately outline the tumor. In particular, imprecise outlining of the tumor might influence the morphological features that quantify the shape and contour of the tumor. Moreover, the texture features, which are extracted from the outlined tumor region, might also be affected by tumor segmentation errors. In this study, tumor outlining was performed by a radiologist with more than thirteen years of experience. Such manual outlining by an experienced operator has been employed in several previous studies, such as [10, 15]. In fact, the manual outlining of the tumor is a time consuming task and its accuracy is subject to the experience level of the radiologist. The future direction of this work is to employ automatic tumor segmentation algorithms, such as [45], that employ advanced image processing techniques to achieve accurate and objective outlining of the tumors.

The multiple-ROI texture analysis approach employed in this study can be extended to reduce the effect of tumor outlining errors. In particular, for each ROI inside the computer-drawn outline, a well-trained classifier can be used to estimate the probability of belonging to the tumor or the surrounding healthy tissue. Such probability estimation can be used to weight the tumor class indicators obtained from the individual ROIs. A customized voting algorithm can be developed to combine the weighted tumor class indicators of the individual ROIs and estimate posterior tumor class likelihood.

4. Conclusion

In this study, an effective approach for BUS image classification is proposed. Texture analysis is carried out by dividing the tumor into a set of nonoverlapping ROIs and processing each ROI individually to estimate its tumor class indicator. The tumor class indicators of all ROIs inside the tumor are combined using a majority voting mechanism to estimate the posterior tumor class likelihood. In addition to the multiple-ROI texture analysis, morphological analysis is used to estimate the posterior tumor class likelihood. A probabilistic approach is employed to fuse the posterior tumor class likelihoods obtained using the texture and morphological analyses. The proposed approach has been employed to classify 110 BUS images. The classification results indicate that the proposed approach achieved classification performance that outperforms conventional texture and morphological analyses. In particular, fusing the multiple-ROI texture analysis and morphological analysis enabled classification accuracy of 98.2%, specificity of 98.4%, and sensitivity of 97.8%. These results suggest that the proposed

approach has the potential to provide the radiologists with an accurate second opinion to reduce the rate of expendable biopsy and minimize BUS image misdiagnosis.

Competing Interests

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

Acknowledgments

This research is supported by the Scientific Research Support Fund (Project Number ICT/2/07/2013), Jordan. The authors would like to thank Dr. Adnan Zayadeen from the Royal Medical Services, Jordan, for verifying the manual outlining of the tumors.

References

[1] J. Ferlay, I. Soerjomataram, R. Dikshit et al., "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer*, vol. 136, no. 5, pp. E359–E386, 2015.

[2] K. Kaul and F. M. Daguilh, "Early detection of breast cancer: is mammography enough?" *Hospital Physician*, vol. 9, pp. 49–55, 2002.

[3] D.-R. Chen and Y.-H. Hsiao, "Computer-aided diagnosis in breast ultrasound," *Journal of Medical Ultrasound*, vol. 16, no. 1, pp. 46–56, 2008.

[4] M. Nothacker, V. Duda, M. Hahn et al., "Early detection of breast cancer: benefits and risks of supplemental breast ultrasound in asymptomatic women with mammographically dense breast tissue. A systematic review," *BMC Cancer*, vol. 9, article 335, 2009.

[5] A. Jalalian, S. B. T. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical Imaging*, vol. 37, no. 3, pp. 420–426, 2013.

[6] H. D. Cheng, J. Shan, W. Ju, Y. H. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: a survey," *Pattern Recognition*, vol. 43, no. 1, pp. 299–317, 2010.

[7] W.-J. Wu, S.-W. Lin, and W. K. Moon, "An artificial immune system-based support vector machine approach for classifying ultrasound breast tumor images," *Journal of Digital Imaging*, vol. 28, no. 5, pp. 576–585, 2015.

[8] C.-M. Lin, Y.-L. Hou, T.-Y. Chen, and K.-H. Chen, "Breast nodules computer-aided diagnostic system design using fuzzy cerebellar model neural networks," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 3, pp. 693–699, 2014.

[9] W. G. Flores, W. C. D. A. Pereira, and A. F. C. Infantosi, "Improving classification performance of breast lesions on ultrasonography," *Pattern Recognition*, vol. 48, no. 4, pp. 1121–1136, 2015.

[10] C.-Y. Chen, H.-J. Chiou, S.-Y. Chou et al., "Computer-aided diagnosis of soft-tissue tumors using sonographic morphologic and texture features," *Academic Radiology*, vol. 16, no. 12, pp. 1531–1538, 2009.

[11] D.-R. Chen, C.-L. Chien, and Y.-F. Kuo, "Computer-aided assessment of tumor grade for breast cancer in ultrasound images," *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 914091, 6 pages, 2015.

[12] W. Gómez, W. C. A. Pereira, and A. F. C. Infantosi, "Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound," *IEEE Transactions on Medical Imaging*, vol. 31, no. 10, pp. 1889–1899, 2012.

[13] W.-J. Wu, S.-W. Lin, and W. K. Moon, "Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images," *Computerized Medical Imaging and Graphics*, vol. 36, no. 8, pp. 627–633, 2012.

[14] W. K. Moon, C.-M. Lo, J. M. Chang, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1091–1098, 2013.

[15] S. K. Alam, E. J. Feleppa, M. Rondeau, A. Kalisz, and B. S. Garra, "Ultrasonic multi-feature analysis procedure for computer-aided diagnosis of solid breast lesions," *Ultrasonic Imaging*, vol. 33, no. 1, pp. 17–38, 2011.

[16] W.-J. Wu and W. K. Moon, "Ultrasound breast tumor image computer-aided diagnosis with texture and morphological features," *Academic Radiology*, vol. 15, no. 7, pp. 873–880, 2008.

[17] R.-F. Chang, W.-J. Wu, W. K. Moon, and D.-R. Chen, "Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors," *Breast Cancer Research and Treatment*, vol. 89, no. 2, pp. 179–185, 2005.

[18] K. Nie, J.-H. Chen, H. J. Yu, Y. Chu, O. Nalcioglu, and M.-Y. Su, "Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI," *Academic Radiology*, vol. 15, no. 12, pp. 1513–1525, 2008.

[19] A. V. Alvarenga, A. F. C. Infantosi, W. C. A. Pereira, and C. M. Azevedo, "Assessing the performance of morphological parameters in distinguishing breast tumors on ultrasound images," *Medical Engineering and Physics*, vol. 32, no. 1, pp. 49–56, 2010.

[20] W. K. Moon, C.-M. Lo, N. Cho et al., "Computer-aided diagnosis of breast masses using quantified BI-RADS findings," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 84–92, 2013.

[21] W.-C. Shen, R.-F. Chang, W. K. Moon, Y.-H. Chou, and C.-S. Huang, "Breast ultrasound computer-aided diagnosis using BI-RADS features," *Academic Radiology*, vol. 14, no. 8, pp. 928–939, 2007.

[22] R.-F. Chang, W.-J. Wu, W. K. Moon, and D.-R. Chen, "Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture analysis," *Ultrasound in Medicine and Biology*, vol. 29, no. 5, pp. 679–686, 2003.

[23] D.-R. Chen, R.-F. Chang, C.-J. Chen et al., "Classification of breast ultrasound images using fractal feature," *Clinical Imaging*, vol. 29, no. 4, pp. 235–245, 2005.

[24] M.-C. Yang, W. K. Moon, Y.-C. F. Wang et al., "Robust texture analysis using multi-resolution gray-scale invariant features for breast sonographic tumor diagnosis," *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2262–2273, 2013.

[25] N. Piliouras, I. Kalatzis, N. Dimitropoulos, and D. Cavouras, "Development of the cubic least squares mapping linear-kernel support vector machine classifier for improving the characterization of breast lesions on ultrasound," *Computerized Medical Imaging and Graphics*, vol. 28, no. 5, pp. 247–255, 2004.

[26] Y.-L. Huang, K.-L. Wang, and D.-R. Chen, "Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines," *Neural Computing and Applications*, vol. 15, no. 2, pp. 164–169, 2006.

[27] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, pp. 610–621, 1973.

[28] D.-R. Chen, Y.-L. Huang, and S.-H. Lin, "Computer-aided diagnosis with textural features for breast lesions in sonograms," *Computerized Medical Imaging and Graphics*, vol. 35, no. 3, pp. 220–226, 2011.

[29] R. M. Rangayyan, N. R. Mudigonda, and J. E. L. Desautels, "Boundary modelling and shape analysis methods for classification of mammographic masses," *Medical and Biological Engineering and Computing*, vol. 38, no. 5, pp. 487–496, 2000.

[30] L.-K. Soh and C. Tsatsoulis, "Texture analysis of sar sea ice imagery using gray level co-occurrence matrices," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 2, pp. 780–795, 1999.

[31] D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Canadian Journal of Remote Sensing*, vol. 28, no. 1, pp. 45–62, 2002.

[32] W. K. Moon, Y.-S. Huang, C.-M. Lo et al., "Computer-aided diagnosis for distinguishing between triple-negative breast cancer and fibroadenomas based on ultrasound texture features," *Medical Physics*, vol. 42, no. 6, pp. 3024–3035, 2015.

[33] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, Burlington, Mass, USA, 4th edition, 2009.

[34] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[35] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[36] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *The Journal of Machine Learning Research*, vol. 6, pp. 1889–1918, 2005.

[37] Y. Xie, S. Wei, X. Wang, S. Xie, and C. Yang, "A new prediction model based on the leaching rate kinetics in the alumina digestion process," *Hydrometallurgy*, vol. 164, pp. 7–14, 2016.

[38] F. M. J. Valckx and J. M. Thijssen, "Characterization of echographic image texture by cooccurrence matrix parameters," *Ultrasound in Medicine and Biology*, vol. 23, no. 4, pp. 559–571, 1997.

[39] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, vol. 10, pp. 61–74, MIT Press, 1999.

[40] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, 2004.

[41] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2012.

[42] E. J. Feleppa, J. Mamou, C. R. Porter, and J. Machi, "Quantitative ultrasound in cancer imaging," *Seminars in Oncology*, vol. 38, no. 1, pp. 136–150, 2011.

[43] M. L. Oelze, W. D. O'Brien Jr., J. P. Blue, and J. F. Zachary, "Differentiation and characterization of rat mammary fibroadenomas and 4T1 mouse carcinomas using quantitative ultrasound imaging," *IEEE Transactions on Medical Imaging*, vol. 23, no. 6, pp. 764–771, 2004.

[44] N. Uniyal, H. Eskandari, P. Abolmaesumi et al., "Ultrasound RF time series for classification of breast lesions," *IEEE Transactions on Medical Imaging*, vol. 34, no. 2, pp. 652–661, 2015.

[45] M. I. Daoud, A. A. Atallah, F. Awwad, and M. Al-Najar, "Accurate and fully automatic segmentation of breast ultrasound images by combining image boundary and region information," in *Proceedings of the IEEE 13th International Symposium on Biomedical Imaging (ISBI '16)*, pp. 718–721, IEEE, Prague, Czech Republic, April 2016.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 6215085, 7 pages
http://dx.doi.org/10.1155/2016/6215085

Research Article

Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images

Wei Li,[1,2] Peng Cao,[1,2] Dazhe Zhao,[1,2] and Junbo Wang[3]

[1]*Medical Image Computing Laboratory of Ministry of Education, Northeastern University, Shenyang 110819, China*
[2]*College of Computer Science and Engineering, Northeastern University, Shenyang 110819, China*
[3]*Neusoft Research Institute, Neusoft Corporation, Shenyang 110179, China*

Correspondence should be addressed to Peng Cao; caopeng@cse.neu.edu.cn

Received 4 August 2016; Revised 4 November 2016; Accepted 22 November 2016

Academic Editor: Kenji Suzuki

Computer aided detection (CAD) systems can assist radiologists by offering a second opinion on early diagnosis of lung cancer. Classification and feature representation play critical roles in false-positive reduction (FPR) in lung nodule CAD. We design a deep convolutional neural networks method for nodule classification, which has an advantage of autolearning representation and strong generalization ability. A specified network structure for nodule images is proposed to solve the recognition of three types of nodules, that is, solid, semisolid, and ground glass opacity (GGO). Deep convolutional neural networks are trained by 62,492 regions-of-interest (ROIs) samples including 40,772 nodules and 21,720 nonnodules from the Lung Image Database Consortium (LIDC) database. Experimental results demonstrate the effectiveness of the proposed method in terms of sensitivity and overall accuracy and that it consistently outperforms the competing methods.

1. Introduction

Lung cancer is becoming one of the main threats to human health at present in the world. The number of deaths caused due to lung cancer is more than prostate, colon, and breast cancers [1]. Early detection of solitary pulmonary nodules (SPNs) is an important clinical indication for early-stage lung cancer diagnosis because SPNs have high probabilities to become malignant nodules [2, 3]. SPNs refer to lung tissue abnormalities that are roughly spherical with round opacity and a diameter of up to 30 mm.

It is therefore an important task to develop computer aided detection (CAD) systems that can aid/enhance radiologist workflow and potentially reduce false-negative findings. CAD is a scheme that automatically detects suspicious lesions (i.e., nodule, polyps, and masses) in medical images of certain body parts and provides their locations to radiologists [4–6]. CAD has become one of the major research topics in medical imaging and diagnostic radiology and has been applied to various medical imaging modalities including computed tomography (CT) [7], magnetic resonance imaging (MRI) [8], and ultrasound imaging [9]. Generally, typical CAD systems for cancer detection and diagnosis (i.e., breast, lung, and polyp) cover four stages as depicted in Figure 1(a), including candidate nodule ROI (Region of Interest) detection, feature extraction, and nodule classification. The stages of feature extraction and nodule classification belong to the false-positive reduction step. Current CAD schemes for nodule characterization have achieved high sensitivity levels and would be able to improve radiologists' performance in the characterization of nodules in thin-section CT, whereas current schemes for nodule detection appear to report many false positives. It is because detection algorithms have high sensitivity that some nonnodule structures (e.g., blood vessels) are labeled as nodules inevitably in the initial nodule identification step. Since the radiologists must examine each identified object, it is highly desirable to eliminate these false positives (FPs) as much as possible while retaining the true positives (TPs). Therefore, significant effort is needed in order to improve the performance levels of current CAD schemes for nodule detection in thin-section CT.

The purpose of false-positive reduction is to remove these false positives (FPs) as much as possible while retaining a relatively high sensitivity [10, 11]. It is a binary classification

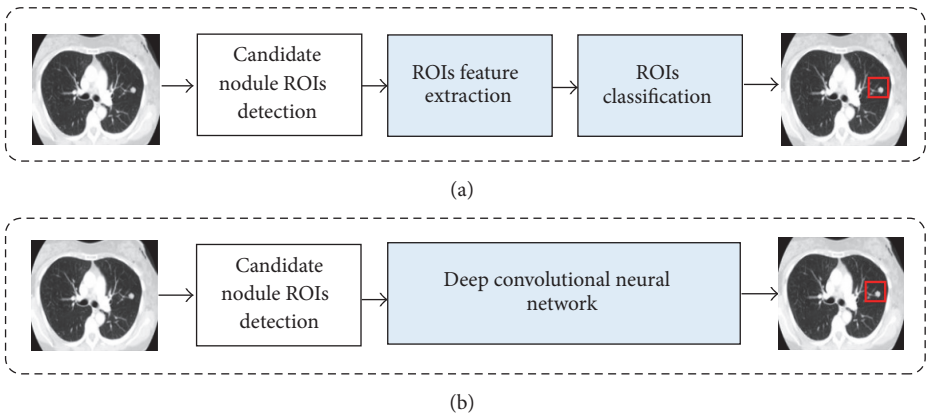


Figure 1: The main components in a general CAD system (a) and the main components in our work (b).

between the nodule and nonnodule, aiming to develop new methods in order to accurately distinguish suspicious regions, leading to significant reduction of FPs with machine learning techniques. The false-positive reduction step, or classification step, the aim of which is to learn a system capable of the prediction of the unknown output class of a previously unseen suspicious nodule with a good generalization ability, is a critical part in the lung nodule detection system. Classification plays an important role in the reduction of false positives in lung computer aided detection and diagnosis methods. Deep learning can be used for both classification and feature learning in various fields such as computer vision and speech. In our work, a deep convolutional neural network is proposed for pulmonary nodule classification using the LIDC database. The method used in CAD system replaces the two components of feature extraction and classification. The input of deep convolutional neural networks in this work is ROI pixel data directly without feature extraction and selection. Compared with the traditional methods, the approach in our work has four advantages as follows.

(i) The representation of nodule ROI is critical for discrimination between true nodule and false nodule. However, it is difficult to obtain good feature representations by human efforts. Our method can learn a good feature representation of ROI without feature extraction and selection.

(ii) Our method takes advantage of the relationships between the internal region and external region of ROI, so as to learn more discriminative knowledge for false-positive reduction.

(iii) Our method can be executed based on the center of the ROI rather than the whole ROI region. Therefore, there is no necessity to obtain the exact margin of the nodules detected in the first step of CAD system.

(iv) The neural networks are trained by large scale ROIs data with nodules and nonnodules more than 60 thousand which are the largest in our knowledge. So the neural network is capable of recognizing a wide range of representations of nodules.

The rest of the paper is organized as follows. Section 2 analyzes the related works. The methodology to recognize nodules is described in Section 3. The experimental results obtained are discussed in Section 4. We conclude this paper in Section 5.

2. Related Work

At present, a lot of works have been done in pulmonary nodule recognition research. The pulmonary nodule recognition involves nodule candidate detection [12] and false-positive reduction [13]. The traditional approaches of false-positive reduction have successive steps: feature extraction [14, 15] and classifier model construction [10, 16]. The most effective features which can be used for classification for lung CT images are, for example, shape, intensity, texture, geometric, gradient, and wavelet. Texture features as Haralick, Gabor, and Local Binary Patterns are used to analyze lung nodules in [17]. MR8, LBP (Local Binary Patterns), Sift descriptor, and MHOG (Multiorientation Histogram of Oriented Gradients) are used for the feature extraction process in [18], and the SURF (Speed-Up Robust Feature) and the LBP descriptors are used to generate the features that describe the texture of common lung nodules in [19]. Mohammad applied an improved LBP feature in lung nodule detection which is robust for noise [20]. Sui et al. used 2D features of circularity, elongation, compactness, moment, and 3D features as surface-area, volume, sphericity, and centroid-offset for lung nodule recognition [21]. Although the feature is well and comprehensively designed, the classifiers in the third step of CAD system still show their deficiencies on classifying the nodule images precisely. Generally speaking, the classifiers are supervised learning approaches in machine learning domain, such as SVM, k-nearest neighbor (k-NN), artificial neural networks (ANNs), and decision tree which have been used in lung nodule classification [22]. In addition, Zhang et al. designed a classifier in a semisupervised way exploring the information from unlabeled images [23]. In order to improve the ensemble classification advantage in lung nodule recognition task, a random forest algorithm with a structure for a hybrid random forest aided by clustering is described in [24].

The imbalance distribution between the amounts of nodule and nonnodule candidates comes out in mostly datasets. Sui et al. present a novel SVM classifier combined with random undersampling and SMOTE for lung nodule recognition [21]. Cao et al. extend the random subspace method to a novel Cost Sensitive Adaptive Random Subspace (CSARS) ensemble to overcome imbalanced data classification [10].

In recent years, deep artificial neural networks have won numerous contests in pattern recognition and machine learning. Convolutional neural networks (CNNs) constitute one such class of models [30]. In 2012, an ensemble CNNs approach achieved the best results on the ImageNet classification benchmark, which is popular in the computer vision community [31]. There has also been popular latest research in area of medical imaging using deep learning with promising results. Suk et al. propose a novel latent and shared feature representation of neuroimaging data of brain using Deep Boltzmann Machine (DBM) for AD/MDC diagnosis [32]. Wu et al. use deep feature learning for deformable registration of brain MR images to improve image registration by using deep features [33]. Xu et al. present the effectiveness of using deep neural networks (DNNs) for feature extraction in medical image analysis as a supervised approach [34]. Kumar et al. propose a CAD system which uses deep features extracted from an autoencoder to classify lung nodules as either malignant or benign on LIDC database, which is similar to our work [35]. Convolutional neural networks have performed better than DBNs by themselves in current literature on benchmark computer vision datasets. The CNNs have attracted considerable interest in machine learning since they have strong representation ability in learning useful features from input data in recent years [36]. Moreover, to the best of our knowledge there has been no work that uses deep convolutional neural networks for lung nodule classification. Therefore, we evaluate the CNN on the computer aided lung nodule.

3. Proposed Method

3.1. Data. The dataset used in this work is the LIDC-IDRI dataset [37], consisting of 1010 thoracic CT scans with nodule size reports and diagnosis reports that serve as a medical imaging research resource. Four radiologists reviewed each scan using two blinded phases. The results of each radiologist's unblinded review were compiled to form the final unblinded review. The LIDC radiologists' annotations include freehand outlines of nodules ≥ 3 mm in diameter on each CT slice in which the nodules are visible, along with the subjective ratings on a five- or six-point scale of the following pathologic features: calcification, internal structure, subtlety, lobulation, margins, sphericity, malignancy, texture, and spiculation. The annotations also include a single mark (an approximate centroid) of nodules ≤ 3 mm in diameter as well as nonnodules ≥ 3 mm.

We included nodules with their annotated centers from the nodule report. The average width and height of the nodule images are 14 pixels, and the median is 12 pixels. The nodules whose sizes are less than $32 * 32$ account for 95.33% of the overall data, and the percentage is 99.991% for less than $64 * 64$ size of nodules.

In the first step of the ROI extraction, the geometric center is computed by the region margin marked in the database. Then region size is determined whether it is larger than $32 * 32$. The $32 * 32$ rectangle region is segmented with the same geometric of the marked region if its size is less than $32 * 32$. Otherwise, a larger size of $64 * 64$ is obtained as a candidate ROI and then is downsampled to $32 * 32$ size finally. There are nonnodule annotated regions extracted by the same way to form the negative sample during the training and testing process. In order to evaluate the effectiveness of the neural networks for different image sizes, dataset is also made with $64 * 64$ size using the same procedure. As a result, a total of 62,492 ROI image patches are extracted from 1,013 LIDC lung image cases containing 40,772 nodules and 21,720 nonnodules.

3.2. Convolutional Neural Network Construction. In computer vision, deep convolutional neural networks (CNNs) have been introduced because they can simulate the behavior of the human vision system and learn hierarchical features, allowing object local invariance and robustness to translation and distortion in the model [36]. CNNs are an alternative type of neural network that can be used to model spatial and temporal correlation while reducing translational variance in signals. The deep convolutional neural networks are built based on the size of input images. The structures of networks are different according to the different image size. A deep CNN proposed in this paper is constructed on $32 * 32$ image ROI data as an example presented in Figure 2.

The convolutional neural networks have two convolutional layers and there is a downsampling layer behind the convolutional layer. Fully connected layers are appended to the last downsampling layer. The first convolutional layer contains 8 feature maps, and the second has 16 ones. The kernel size is $5 * 5$ in all convolutional layers and the step of kernel is 1. The kernel size is $2 * 2$ for all the downsampling layers and the step is 2. The first fully connected layer contains 150 nodes and there are 100 nodes in the second fully connected layer. There are 50 nodes in the third fully connected layer and the last layer only has two nodes which are presented as output probabilities of nodule and nonnodule. The ROI region can be recognized as nodule or nonnodule by the output probabilities. In the same way, the convolutional neural networks can be constructed for $64 * 64$ size input image only and the convolution kernel size, convolution kernels moving step, feature map, and the number of nodes are adjusted which are not discussed here.

3.3. Neural Network Training. The deep CNNs described in above section are trained by the LIDC ROI image set extracted in Section 3.1. Firstly, the random initialization of the network weights is conducted and then ROI images are normalized as input into the neural network. At the training stage, the images entered into the network are with labels; that is, each ROI area is known as pulmonary nodules or not. Given each layer in the network input as X and output as Y, the current layer as the convolutional or fully connected layer is calculated as $Y = \max(0, \omega X + B)$, where ω is the

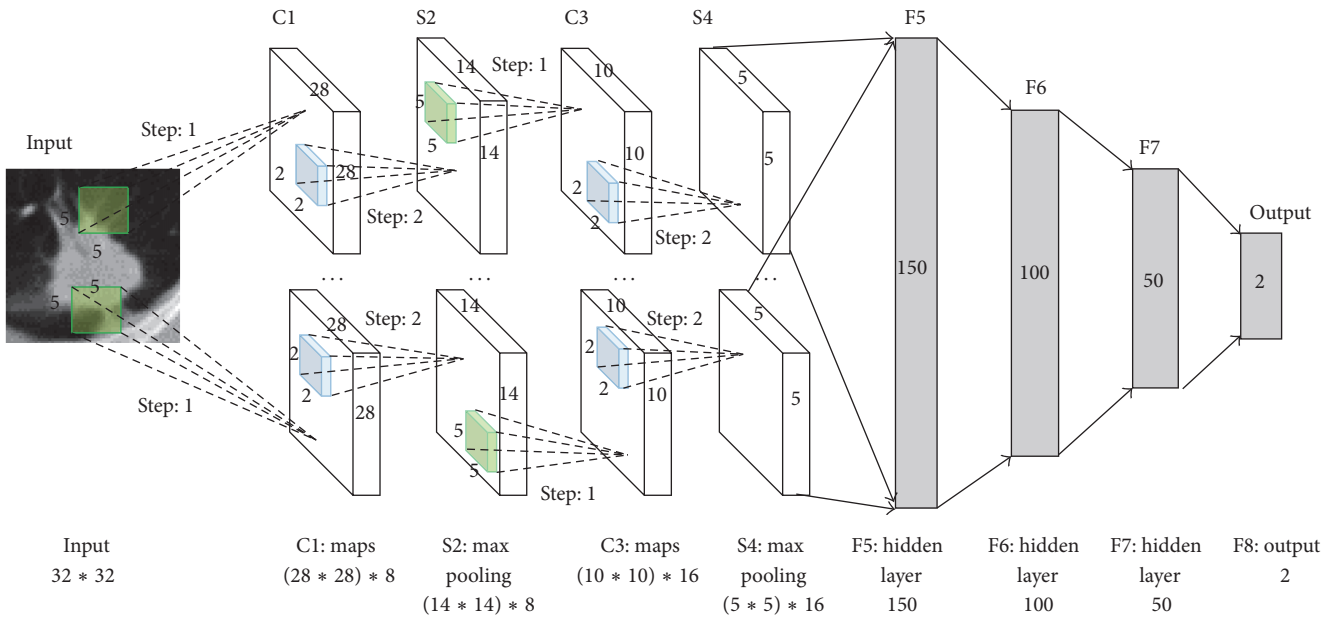


Figure 2: Architecture of our CNN for nodule recognition. The input data is ROI image pixels as a 1024-dimensional vector, and the number of output neurons of the network is 2 (nodule: 1 and nonnodule: 0). The numbers of neurons in the other layers are set to 6272, 1568, 1600, 250, 150, 100, and 50.

current layer weights corresponding to each node and B is the bias parameter. The formulation is $Y = \max(X)$ for the downsampling layers. The output layer is a softmax layer that predicts the probability of the nodule class. Two probabilities are obtained in the output layer after computing operations followed as above descriptions from input image data. The new weights values can be updated by backpropagation algorithm using the two probabilities and the label data with 0 or 1 [16]. The training process is terminated when the accuracy is up to predetermined value or the convergence condition. Finally, the evaluation is conducted on the testing data with the trained model.

4. Experiments

The experimental evaluations are conducted on LIDC database. The test scheme is designed as two different strategies. One is 10-fold cross-validation (CF-test) and the other is that the dataset is divided into the training data (85.7%) and testing data (DD-test). Since all the previous works are based on the manually designed features while the proposed approach in this paper is based on feature learning and nodule recognition by deep convolutional neural networks, it is not possible to directly compare our method with them on the same LIDC dataset. All experiments are conducted on a desktop computer with Intel Core 2 CPU of 2.80 GHz, 8 GB memory, and Windows 7. The algorithm is implemented by C++ in Microsoft Visual Studio 2010. The performance is shown as in Table 1 with both CF-test and DD-test. The tests T1, T2, and T3 are used by the strategy of CF-test and the parameters of convolutional map size and momentum for weight updating are set as 6 and 0.9. The learning rates in T1 and T2 are 0.0005 while T3 is 0.001. The image sizes in T2 and T3 are $64 * 64$ and the other ones are $32 * 32$. The tests T4 and T5 are used by the strategy of DD-test where the momentum and learning rate are set as 0.95 and 0.0005, respectively. However, the convolutional map size is set to 6 for T4 and 8 for T5 test. In CF-test, the learning rate keeps unchanged in the entire training process. However, the learning rates in T4 and T5 tests are decreased by 5/6 of last iteration once the value of precise up to 0.85. Figure 3(a) shows that the performance of accuracy and error trend in CF-test and the same evaluation result is presented in Figure 3(b) which has the maximum iteration to 50.

Table 1: Performance for CF-test and DD-test.

TID	Accuracy	Sensitivity	FP/exam	F-measure	Time (s)
T1	0.855	0.855	4.276	0.870	5,236
T2	0.849	0.866	**3.957**	0.858	28,761
T3	0.857	0.871	4.459	0.864	19,302
T4	**0.864**	**0.890**	5.546	**0.877**	19,993
T5	0.843	0.871	5.540	0.857	21,920

The learning rate is changing in the DD-test benchmark which is shown in Figure 4. In DD-test evaluation, the training process is conducted on the training dataset which will be shuffled at the beginning of training at every iteration, and then the model is applied on the testing dataset which is not changed in the entire testing time. Therefore, a new evaluation result is obtained in each iteration. From Table 1, the deep convolutional neural networks obtain a promising performance on pulmonary nodule recognition on CT images. The best accuracy is 0.864 and sensitivity is 0.890. The results also demonstrate that the larger value of the momentum and learning rate can achieve a fast convergence performance.

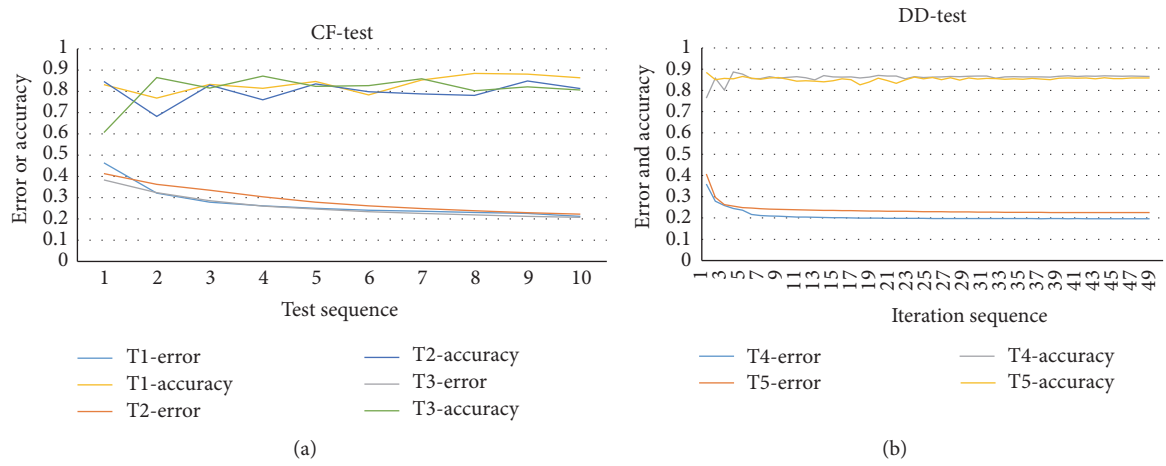


Figure 3: The classification performance with respect to error and accuracy with iteration number.

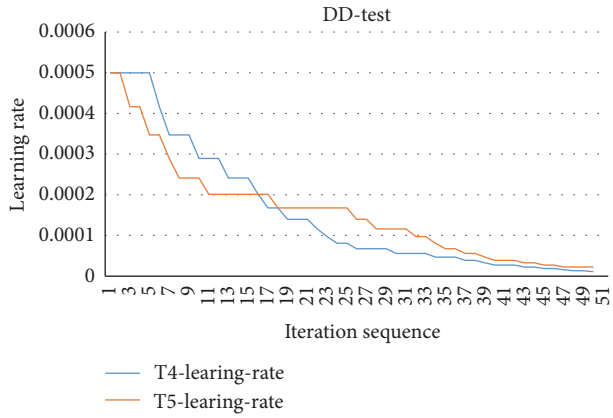


Figure 4: The learning rate changes in training process.

The results shown in Figures 3 and 4 demonstrate that the learning rate converges more smoothly compared with CF-test. Although the change of accuracy is large at the top iteration in CF-test, the error is increasing in training gradually and the whole networks are stable in the last. However, the performance with respect to error and accuracy becomes much more stable after several iterations. This behavior is correlated with the change of learning rate, because when the network obtains an optimal point then the training process gets stable. Overall, the deep convolutional neural network shows its stability and robustness in the training process. Moreover, the CNN framework is effective and efficient in classification.

In order to show the performance of the deep learning based method, we compared it with the state-of-the-art methods designed for lung nodule detection. The result is shown in Table 2. Strictly speaking, it is hard to compare to other reported works on the lung nodule detection problem. This is because most work does not employ the whole LIDC datasets. From the results in Table 2, our empirical results are very encouraging and have demonstrated the promise of the proposed method in the lung nodule detection with respect to sensitivity and FP/exam.

Table 2: Comparison of studies on nodule detection.

Work	Database	Cases	Sensitivity (%)	FP/exam
Proposed method	LIDC	1010	87.1	4.622
Netto et al. [25]	LIDC	29	85.9	0.138
Pei et al. [26]	LIDC	30	100	8.4
Pu et al. [27]	LIDC	52	81.5	6.5
Namin et al. [28]	LIDC	63	88.0	10.3
Messay et al. [29]	LIDC	84	82.66	3

5. Conclusions

In this paper, a method of pulmonary nodule recognition using deep convolutional neural networks is presented. The deep convolutional neural network can take advantage of the training dataset to enable the algorithm to automatically select the best representation as the feature representation of the image. Through the training of the training dataset, the approach obtains much more general characteristics of pulmonary nodules and higher accuracy while retaining relatively better robustness. We plan to extend the proposed method to be capable of benign and malignant classification in the future. The algorithm will be accelerated by GPU computing for convolution operation.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This research was supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China under Grant 2014BAI17B01, the National Natural Science Foundation of

China (61502091), the Fundamental Research Funds for the Central Universities under Grant nos. N140403004, N150408001, and N140407001, and the Postdoctoral Science Foundation of China (2015M570254).

References

[1] P. B. Bach, J. N. Mirkin, T. K. Oliver et al., "Benefits and harms of CT screening for lung cancer: a systematic review," *JAMA*, vol. 307, no. 22, pp. 2418–2429, 2012.

[2] H. T. Winer-Muram, "The solitary pulmonary nodule," *Radiology*, vol. 239, no. 1, pp. 34–49, 2006.

[3] K. Suzuki, F. Li, S. Sone, and K. Doi, "Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network," *IEEE Transactions on Medical Imaging*, vol. 24, no. 9, pp. 1138–1150, 2005.

[4] K. Suzuki, "A review of computer-aided diagnosis in thoracic and colonic imaging," *Quantitative Imaging in Medicine and Surgery*, vol. 2, pp. 163–176, 2012.

[5] M. Tan, R. Deklerck, B. Jansen, M. Bister, and J. Cornelis, "A novel computer-aided lung nodule detection system for CT images," *Medical Physics*, vol. 38, no. 10, pp. 5630–5645, 2011.

[6] M. N. Gurcan, B. Sahiner, N. Petrick et al., "Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system," *Medical Physics*, vol. 29, no. 11, pp. 2552–2558, 2002.

[7] S.-H. Peng, D.-H. Kim, S.-L. Lee, and M.-K. Lim, "Texture feature extraction based on a uniformity estimation method for local brightness and structure in chest CT images," *Computers in Biology and Medicine*, vol. 40, no. 11-12, pp. 931–942, 2010.

[8] Y. Zhang, S. Wang, P. Phillips, Z. Dong, G. Ji, and J. Yang, "Detection of Alzheimer's disease and mild cognitive impairment based on structural volumetric MR images using 3D-DWT and WTA-KSVM trained by PSOTVAC," *Biomedical Signal Processing and Control*, vol. 21, pp. 58–73, 2015.

[9] C.-Y. Chang, S.-J. Chen, and M.-F. Tsai, "Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images," *Pattern Recognition*, vol. 43, no. 10, pp. 3494–3506, 2010.

[10] P. Cao, D. Zhao, and O. Zaiane, "Cost sensitive adaptive random subspace ensemble for computer-aided nodule detection," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS '13)*, pp. 173–178, Porto, Portugal, June 2013.

[11] K. Suzuki, J. Shiraishi, H. Abe, H. MacMahon, and K. Doi, "False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network," *Academic Radiology*, vol. 12, no. 2, pp. 191–201, 2005.

[12] K. Suzuki, "A supervised 'lesion-enhancement' filter by use of a massive-training artificial neural network (MTANN) in computer-aided diagnosis (CAD)," *Physics in Medicine and Biology*, vol. 54, no. 18, pp. S31–S45, 2009.

[13] K. Suzuki, S. G. Armato III, F. Li, S. Sone, and K. Doi, "Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography," *Medical Physics*, vol. 30, no. 7, pp. 1602–1617, 2003.

[14] X. Ye, X. Lin, J. Dehmeshki, G. Slabaugh, and G. Beddoe, "Shape-based computer-aided detection of lung nodules in thoracic CT images," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 7, pp. 1810–1820, 2009.

[15] D. S. Paik, C. F. Beaulieu, G. D. Rubin et al., "Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT," *IEEE Transactions on Medical Imaging*, vol. 23, no. 6, pp. 661–675, 2004.

[16] T. Sun, J. Wang, X. Li et al., "Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 2, pp. 519–524, 2013.

[17] F. Han, H. Wang, G. Zhang et al., "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *Journal of Digital Imaging*, vol. 28, no. 1, pp. 99–115, 2014.

[18] F. Zhang, Y. Song, W. Cai et al., "Lung nodule classification with multilevel patch-based context analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1155–1166, 2014.

[19] F. Amal, A. Asem, G. James et al., "Feature-based lung nodule classification," in *Advances in Visual Computing: 6th International Symposium, ISVC 2010, Las Vegas, NV, USA, November 29– December 1, 2010, Proceedings, Part III*, vol. 6455, pp. 79–88, Springer, Berlin, Germany, 2010.

[20] H. S. Mohammad, "Lung nodule detection based on noise robust local binary pattern," *International Journal of Scientific and Engineering Research*, vol. 5, pp. 356–362, 2014.

[21] Y. Sui, Y. Wei, and D. Zhao, "Computer-aided lung nodule recognition by SVM classifier based on combination of random undersampling and SMOTE," *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 368674, 13 pages, 2015.

[22] Y. Song, W. Cai, Y. Zhou, and D. D. Feng, "Feature-based image patch approximation for lung tissue classification," *IEEE Transactions on Medical Imaging*, vol. 32, no. 4, pp. 797–808, 2013.

[23] F. Zhang, Y. Song, W. Cai et al., "A ranking-based lung nodule image classification method using unlabeled image knowledge," in *Proceedings of the IEEE 11th International Symposium on Biomedical Imaging (ISBI '14)*, pp. 1356–1359, Beijing, China, May 2014.

[24] S. L. A. Lee, A. Z. Kouzani, and E. J. Hu, "Random forest based lung nodule classification aided by clustering," *Computerized Medical Imaging and Graphics*, vol. 34, no. 7, pp. 535–542, 2010.

[25] S. M. B. Netto, A. C. Silva, R. A. Nunes, and M. Gattass, "Automatic segmentation of lung nodules with growing neural gas and support vector machine," *Computers in Biology and Medicine*, vol. 42, no. 11, pp. 1110–1121, 2012.

[26] X. Pei, H. Guo, and J. Dai, "Computerized detection of lung nodules in CT images by use of multiscale filters and geometrical constraint region growing," in *Proceedings of the 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE '10)*, pp. 1–4, Chengdu, China, June 2010.

[27] J. Pu, B. Zheng, J. K. Leader, X.-H. Wang, and D. Gur, "An automated CT based lung nodule detection scheme using geometric analysis of signed distance field," *Medical Physics*, vol. 35, no. 8, pp. 3453–3461, 2008.

[28] S. T. Namin, H. A. Moghaddam, R. Jafari, M. Esmaeil-Zadeh, and M. Gity, "Automated detection and classification of pulmonary nodules in 3D thoracic CT images," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '10)*, pp. 3774–3779, Istanbul, Turkey, October 2010.

[29] T. Messay, R. C. Hardie, and S. K. Rogers, "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery," *Medical Image Analysis*, vol. 14, no. 3, pp. 390–406, 2010.

[30] A. Alpher and J. P. N. Fotheringham-Smy, "Convolutional networks and applications in vision," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '10)*, pp. 253–256, Paris, France, 2010.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '12)*, pp. 1–4, Lake Tahoe, Nev, USA, 2012.

[32] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[33] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of MR brain images," *Medical Image Computing and Computer-Assisted Intervention*, vol. 16, part 2, pp. 649–656, 2013.

[34] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I.-C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '14)*, pp. 1626–1630, Shanghai, China, May 2014.

[35] D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in CT images," in *Proceedings of the 12th Conference on Computer and Robot Vision (CRV '15)*, pp. 133–138, IEEE, Halifax, Canada, June 2015.

[36] L. Yann and B. Yoshua, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10, 1995.

[37] S. G. Armato III, G. McLennan, L. Bidaut et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 1091279, 7 pages
http://dx.doi.org/10.1155/2016/1091279

Research Article

Lung Nodule Image Classification Based on Local Difference Pattern and Combined Classifier

Keming Mao and Zhuofu Deng

College of Software, Northeastern University, Shenyang, Liaoning Province 110004, China

Correspondence should be addressed to Keming Mao; maokm@mail.neu.edu.cn

Received 5 September 2016; Accepted 6 November 2016

Academic Editor: Ayman El-Baz

This paper proposes a novel lung nodule classification method for low-dose CT images. The method includes two stages. First, Local Difference Pattern (LDP) is proposed to encode the feature representation, which is extracted by comparing intensity difference along circular regions centered at the lung nodule. Then, the single-center classifier is trained based on LDP. Due to the diversity of feature distribution for different class, the training images are further clustered into multiple cores and the multicenter classifier is constructed. The two classifiers are combined to make the final decision. Experimental results on public dataset show the superior performance of LDP and the combined classifier.

1. Introduction

Lung cancer is among the most common medical conditions worldwide, and it is responsible for 1.56 million deaths as of the year 2012 [1]. Overall, 16.8% of people in the United States that are diagnosed with lung cancer survive for five years after the diagnosis, while its outcomes on average are worse in the developing countries [2]. It is showed that using low-dose computed tomography (CT) for early detection can significantly reduce the mortality of lung cancer [3]. Therefore, as a result, there is urgent desire for lung nodule CT image analysis in an efficient and convenient way.

Usually, a lung nodule is characterized by its bright appearance compared with its surrounding regions. Commonly, lung nodules can be classified into four different types according to their relative locations with neighbor pulmonary structures [4]. Here (A), (B), (C), and (D) are used to denote four types of lung nodule:

(A) Well-circumscribed nodule: without any connection to other pulmonary structures

(B) Juxtavascular nodule: with uncertain connections to surrounding vessels

(C) Pleural-tail nodule: with a thin connection between the nodule and the pleural

(D) Juxtapleural nodule: with a large proportional connection between the nodule and the pleural

Demonstrations of four types of lung nodule images are shown in Figures 1(a)–1(d), respectively. The analysis of nodule morphology is a crucial step in the assessment of nodule malignancy [5]. Traditionally, this work is done by the expert manually. It is highly affected by his competence and status, and the efficiency is inevitably weakened for its time consuming. Therefore, automatic lung nodule type classification using computer vision technology is necessary to provide a supplementary medical treatment for the physician. The aim of this work is to automatically classify lung nodule CT image patches into four types with high performance.

Generally, medical image classification contains two main steps: (1) feature extraction and representation and (2) classifier construction. In the first stage, medical image is expressed with high dimensional feature vector, which denotes the texture, color, orientation, and so forth. In the second stage, supervised or unsupervised based learning methods are used to construct the classifier given the labeled training dataset. As a hot study area, there has been a lot of research on lung node image classification. Ciompi et al. focus on designing a descriptor which samples intensity profiles along circular patterns [5], and then a spectrum is computed by Fourier transform. The spectrum is clustered to form a library, and

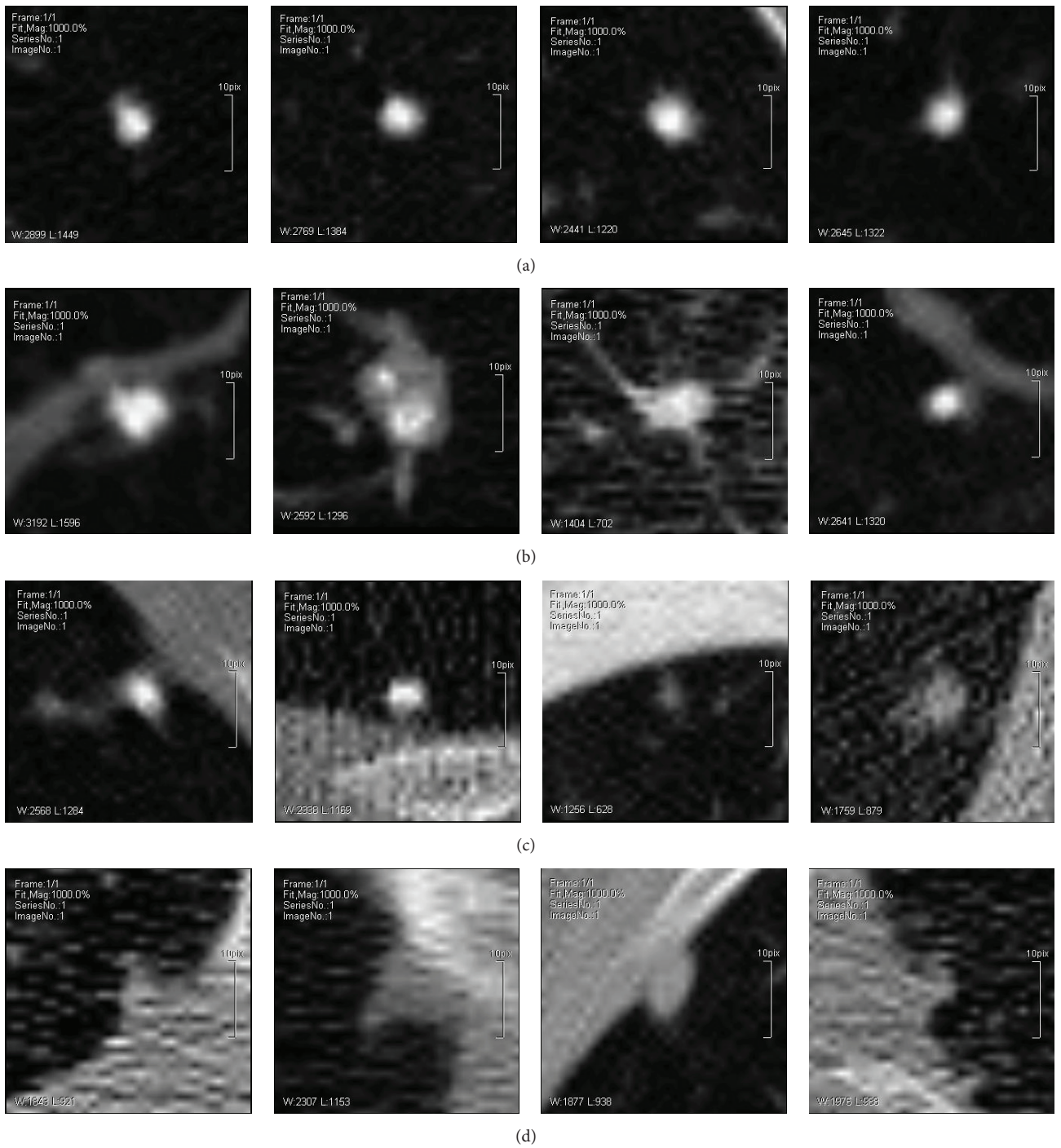


Figure 1: Sample images from the four types with (A), (B), (C), and (D) from left to right, respectively.

bag of frequency is used to construct the feature vector. Song et al. use the region-based energy method to label the background and foreground [6]. The locations of lung nodules with respect to the other structures are gained, and this information is used to construct the feature vector. Farag et al. first applied SIFT descriptor, and PCA and LDA are used for dimension reduction. Then, an adopted Daugman Iris Recognition algorithm is implemented and complex Gabor response is obtained [7]. Zhang et al. first used traditional supervised learning method to construct a bipartite graph [8]. The relationship between test image and training images is used to construct the ranking score and contribution score, and the final classification result is gained. Jacobs et al. propose a segmented-based method [9]. It characterizes the nodule as solid, part-solid, and nonsolid and then a supervised learning method is implemented. In another method, shape features such as smoothness and irregularity of a nodule are used to construct the feature representation

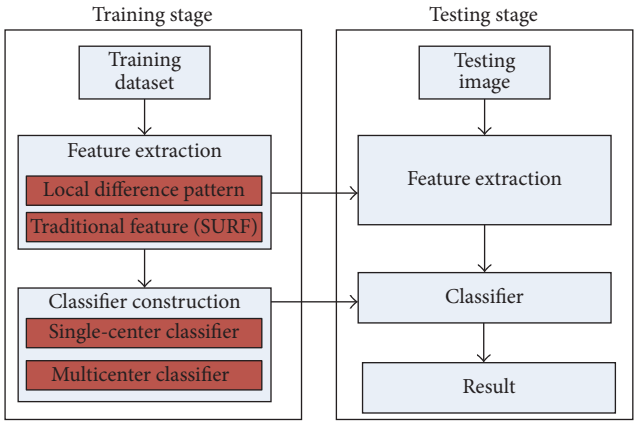


Figure 2: Framework of the proposed method.

[10]. Samala et al. use nine-feature descriptors for lung nodule representation which were often used by radiologists [11]. Lowe uses SIFT representation to characterize the feature of nodule, and then LDA is used to construct the classifier [12]. Maldonado et al. proposes a method that nodule patches are clustered to construct the feature dictionary, and then the testing nodule voxels are labeled [13]. Song et al. first clustered images to a sparse representation based on spectral analysis, and test image is formed with sparse representation. Finally, classifier is constructed by a fusing method [14]. Zhang et al. use a supervised learning method to find four probability values that belongs to each type [15]. Then, a weighed Clique Percolation method is implemented to discover the overlapping of lung nodules that belong to different type.

There are many methods about lung nodule image classification. However, the complex structure of the medical image causes the classification high variance intraclass and high similarity interclass. Therefore, the automatic medical image classification is still a challenging problem. Most of the existing methods adopt generic feature representations which is commonly used in computer vision domain. These methods lack specialized analysis for the texture and shape of lung nodule. On the other hand, using one classifier scheme, whether supervised based or unsupervised based may not be well matched with the lung nodule classification. Facing the above mentioned problems, this paper proposes a novel lung nodule representation and image classification method. As shown in Figure 2, the training stage learns the classification model, and the model is used in testing stage for image classification. In feature extraction step, a novel feature Local Difference Pattern (LDP for abbreviation) is designed based on the gray level difference between lung nodule and its neighbors region. The LDP representation is more specialized and comprehensive. In the step of classifier construction, single-center classifier is first constructed using supervised learning and LDP feature representation. In the next step, labeled images are clustered into multiple centers using unsupervised learning method. The multicenter classifier is then constructed based on the basis of the similarity between the testing image and multiple centers. These two classifiers are combined to construct the final classifier. In testing stage, the image is represented as the same scheme in training stage, and the classification result can be gained using the final classifier. The main contributions of the paper are as follows:

(i) First, based on the analysis of the characteristic of lung nodule and the distribution of the corresponding tissues, a novel feature representation, LDP, is proposed. The new feature is suited for reflect the distinguishing feature of different types of lung nodule.

(ii) Second, generative model and discriminative model are used to construct single-center and multicenter classifiers. These two classifiers complement each other, which makes the classification more robust.

The structure of this paper is organized as follows. Local Difference Pattern is given in Section 2. Classifier construction is given in Section 3. Experimental results are shown in Section 4. Section 5 concludes this paper.

2. Local Difference Pattern

As shown in Figure 1, different types of lung nodule can be characterized by various features, while the size and gray level of the nodule itself could vary to a certain distance. So, this paper focuses on extracting the feature that reflects the gray level difference between the nodule and its neighbor regions.

This paper proposes the Local Difference Pattern (LDP) to describe the local feature of lung nodule image. As shown in Figures 3(a)–3(d) give four types of lung nodule image, and each has three concentric circles with the nodule in the center circle. LDP is extracted according to the concentric circle regions. Figure 3(e) gives the detailed information of subregion partition used for feature extraction. The center circle is denoted as C, and the out layer circles are divided into four parts according to four quadrants. r_i^j is the average gray level of the corresponding region, where superscript j means the number of circle and subscript i means the number of quadrant.

Moreover, one of the most important objectives is the rotation invariant of the local feature. Before LDP extraction, some adjustment should be done to the original image patches. By the aid of design mode from other local feature, that is, SIFT, SURF, and so forth [16], the main direction of the lung nodule image is calculated first, and then LDP can be extracted in the rotated image according to the main direction, as shown in Figure 4. For the lung nodule images are collected with the same resolution, so the scale of the feature cannot be considered here.

In the light of the above description, LDP is defined as follows:

$$\begin{aligned}\mathrm{LDP}(I) = \Big\{ & C, r_i^1, r_i^2, \mathrm{sign}\left(r_i^1 - C\right), \mathrm{sign}\left(r_i^2 - C\right), \\ & \mathrm{sign}\left(r_i^1 - r_{(i+1) \bmod 4}^1\right), \mathrm{sign}\left(r_i^2 - r_{(i+1) \bmod 4}^2\right), \\ & \mathrm{sign}\left(r_i^1 - r_i^2\right)\Big\}, \quad (1 \le i \le 4).\end{aligned} \tag{1}$$

As shown in (1), LDP(I) means feature vector of lung nodule image I, which is composed of multidimensional data.

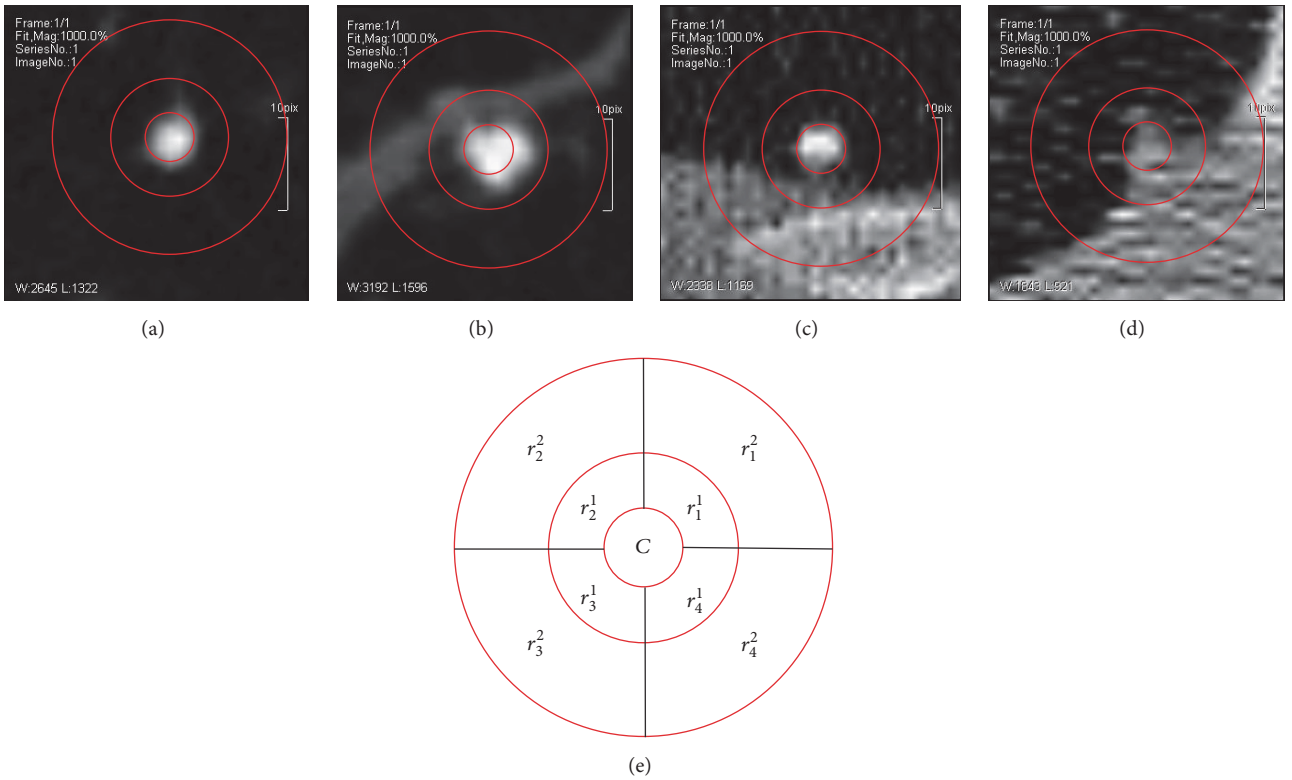


FIGURE 3: Demonstration of Local Difference Pattern. (a)–(d) are four types of lung nodule images, red circles denote the region used for feature extraction. (e) denotes the detail region partition used for feature extraction.

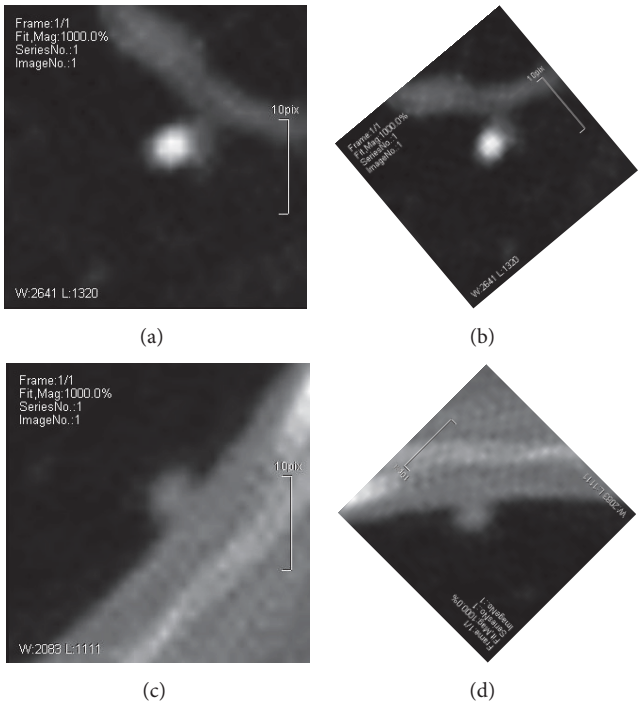


FIGURE 4: Rotation of the lung nodule images. (a) and (c) are the traditional images, while (b) and (d) are their rotated images according to the main direction, respectively.

$\text{sign}(r_i^1 - C)$ and $\text{sign}(r_i^2 - C)$ denote the gray level difference between the center and the outlier circles. $\text{sign}(r_i^1 - r_i^2)$ denotes the gray level difference between 1th and 2nd circle in different quadrant. $\text{sign}(r_i^1 - r_{(i+1) \bmod 4}^1)$ and $\text{sign}(r_i^2 - r_{(i+1) \bmod 4}^2)$ denote the gray level difference between neighbor quadrants in a counterclockwise direction inside one concentric circle. Totally, a 29-dimensional feature vector is used to represent the LDP.

3. Classifier Construction

In this section, single-center classifier and multicenter classifier are constructed, respectively, and a combined one is further build. Illustrations are given in detail as follows.

3.1. Single-Center Classifier. Given the labeled image dataset, LDP feature is first extracted for each training lung nodule image, and then a supervised learning method is used straightly. Here, linear SVM is adopted to construct the classifier model, and it is called single-center classifier f^S. The classifier f^S outputs the possibility that one image belongs to each type of lung nodule.

3.2. Multicenter Classifier. The lung nodule images are not easy to classify for there exist large intraclass variance and high interclass similarity. And due to the multiple distribution nature of diversity for the image, a single supervised classifier is probably insufficient to catch the diverse representations of one class data. Thus, this paper applies one more step to the algorithm. By implement clustering with 64-dimensional feature vector SURF of each training image, images with the same class label are further clustered to form some centers, which can be represented as follows:

$$C^k = \left\{C_1^k, C_2^k, \ldots, C_n^k\right\}, \quad 1 \le k \le 4, \tag{2}$$

where superscript k means the class label and subscript i denotes multiclusters in one class. n denotes the number of center. Given an image for testing, its probability that belongs to four types of lung nodules can be computed as follows:

$$CF_i^k = \frac{\text{Num}\left(C_i^k\right)}{\text{Num}\,(k)}, \quad 1 \le k \le 4,\ 1 \le i \le 5, \tag{3}$$

$$S = CF_{i1}^1 + CF_{i2}^2 + CF_{i3}^3 + CF_{i4}^4, \tag{4}$$

$$f^M = \left[\frac{CF_{i1}^1}{S}, \frac{CF_{i2}^2}{S}, \frac{CF_{i3}^3}{S}, \frac{CF_{i4}^4}{S}\right]. \tag{5}$$

As shown in (3), $\text{Num}(C_i^k)$ denotes the number of training images in class k which belongs to ith center. $\text{Num}(k)$ denotes the number of training images in class k. CF_i^k denotes the frequency of center i in class k. Given a test image X, let $C_{i1}^1, C_{i2}^2, C_{i3}^3, C_{i4}^4$ be its centers of four lung nodule types; then, (4) and (5) can be used to construct the multicenter classifier f^M, which gives the values of probability that X belongs to each of four types, respectively:

$$F = w * f^S + (1 - w) * f^M. \tag{6}$$

As shown in (6), the single-center classifier and multicenter classifier are combined to get the final classifier F, where w is the weighted parameter.

4. Experimental Evaluation

4.1. Dataset and Program Implementation. In this section, the public available dataset is used for the experiment evaluation [17]. The dataset contains 379 lung nodule images with center position of nodule annotated, which are comprised of 50 distinct CT lung scans. The lung nodules are classified into four types according to the instruction by an expert.

The lung nodule images are cropped from the original CT images according to the position of nodule center. The original CT image is with a resolution of 512 pixel $*$ 512 pixel, and the cropped image patches are too small to implement the computer vision algorithm. Therefore, the cropped images are further interpolated to 160 pixel $*$ 160 pixel with the bicubic method. All the programs are implemented using Matlab 2012 programming language and tested on a Pentium Dual-2.4 CPU, 2 G RAM PC.

4.2. Parameter Setting. L_{r1}, L_{r2}, and L_{r3}, denoted as the size of three concentric circles, along with classifier weight w and the number of multiclusters in each class n are evaluated with comprehensive testing. As shown in Table 1, the option range of L_{r1} is 20–45 pixels, with a step of 5 pixels, the option range of L_{r2} is 70–95 pixels, with a step of 5 pixels, the option range of L_{r3} is 100–125 pixels, with a step of 5 pixels, the option range of w is 0.3–0.8, with a step of 0.1, and the option range of n is 3–7, with a step of 1. So there are 6480 ($6 * 6 * 6 * 6 * 5$) combinations of parameters setting. After complete testing, L_{r1}, L_{r2}, and L_{r3}, the radii of three concentric circles, are set with 35 pixels, 90 pixels, 105 pixels, respectively. The weight of combined classifier w is assigned with 0.6. The number of multiclusters in each class n is set as 5. This set of parameters gives the highest classification rate.

4.3. The Proportion of Training Dataset versus Classification Rate. The proportion of training dataset may have influence on classification rate of the algorithm. In this subsection, training dataset is selected randomly with the proportion from 10% to 90%, with a step of 5%, and the remainder is used for testing. The testing is performed many times and the average classification rate is computed.

Figure 5 gives the demonstration of proportion of training dataset versus classification rate. As can be seen from the figure, the classification rate is raised as the proportion of training dataset is increased. That means more training data can incorporate more information, and a better data representation diversity can be gained, and therefore the performance is enhanced. Meanwhile, when the proportion exceeds some value the classification rate is tend to be stable.

4.4. Average Classification Rate. In order to evaluate the classification rate comparison between different methods, five algorithms are used for testing, which are composed of

Table 1: The values of parameters used in the proposed method.

Notation	Description	Option range	Determined value
L_{r1}	Radius of 1st concentric circle	20–45 pixels (step with 5 pixels)	35 pixels
L_{r2}	Radius of 2nd concentric circle	70–95 pixels (step with 5 pixels)	90 pixels
L_{r3}	Radius of 3rd concentric circle	100–125 pixels (step with 5 pixels)	105 pixels
w	Weight of combined classifier	0.3–0.8 (step with 0.1)	0.6
n	Number of multiclusters	3–7 (step with 1)	5

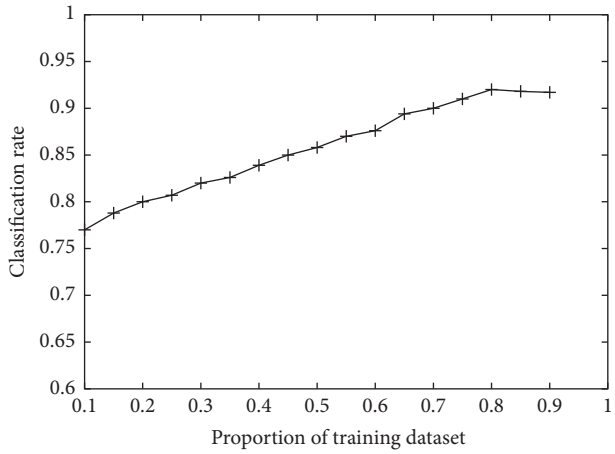


Figure 5: The influence of proportion of training dataset on classification rate.

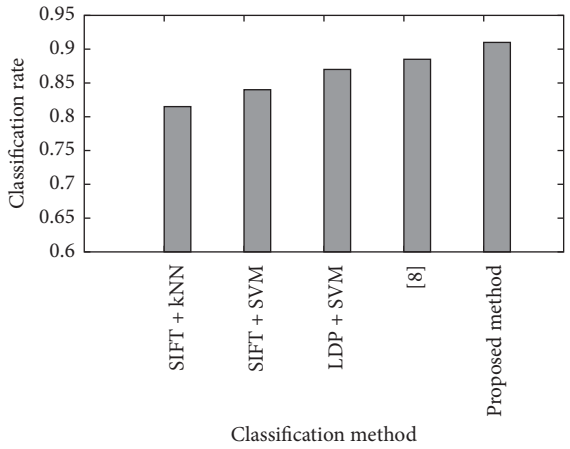


Figure 6: The classification rate among five methods.

various feature representation and classifier. The classification rate is the average value for different training dataset. Figure 6 gives the comparison result. It is shown that LDP + SVM has a higher performance than SIFT + kNN and SIFT + SVM, which means LDP designed in this paper contains more useful information to represent the local feature. Among all five algorithms, the proposed method demonstrates the best performance.

4.5. ROC Testing. ROC curves are a regular tool for illustrating the performance of a classifier system, and the curve can

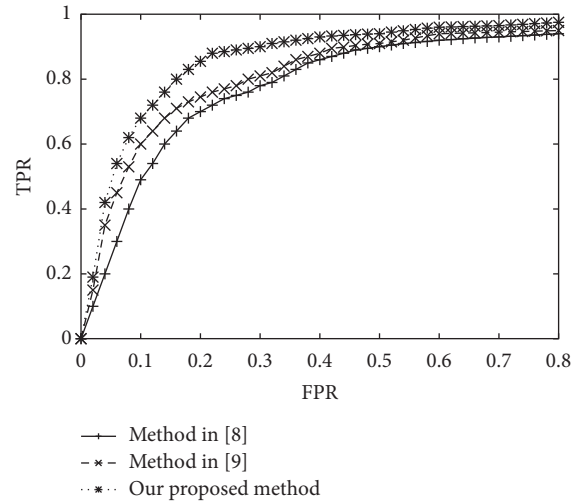


Figure 7: The ROC curve testing with different methods.

be gained by plotting the true positive rate (TPR) against the false positive rate (FPR) at varied discrimination threshold settings. Some recent algorithms are chosen for comparison with our proposed one [8, 9], and the results are given in Figure 7. It can be seen clearly from the demonstration that the proposed method has the superior ROC curves characteristic.

5. Conclusion

This paper proposes a method for lung nodule image classification. First, a novel local feature representation, Local Difference Pattern, is designed, which can catch more information from the lung nodule and its neighbor regions. And a single-center classifier is constructed according to LDP and SVM. Then, a multicenter classifier is designed by clustering the SURF feature of lung nodule image and computing the similarity between testing image and multiple centers. Finally, the two classifiers are combined to implement the classification. The proposed method aims to extract more useful feature and decrease the gap between high variance intraclass and high similarity interclass. Evaluation on public dataset shows that our proposed method outperforms other methods for lung nodule image classification. Our future works will focus on designing more accurate feature representation methods for lung nodule image, such as autoencoder and convolutional neural network.

Competing Interests

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

References

[1] *World Cancer Report 2014*, World Health Organization, Geneva, Switzerland, 2014.

[2] National Cancer Institute, *Surveillance, Epidemiology and End Results Program*, 2014.

[3] D. R. Aberle, A. M. Adams, C. D. Berg et al., "Reduced lung-cancer mortality with low-dose computed tomographic screening," *The New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, 2011.

[4] S. Diciotti, G. Picozzi, M. Falchini, M. Mascalchi, N. Villari, and G. Valli, "3-D segmentation algorithm of small lung nodules in spiral CT images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 1, pp. 7–19, 2008.

[5] F. Ciompi, C. Jacobs, E. T. Scholten et al., "Bag-of-frequencies: a descriptor of pulmonary nodules in computed tomography images," *IEEE Transactions on Medical Imaging*, vol. 34, no. 4, pp. 962–973, 2015.

[6] Y. Song, W. Cai, Y. Wang, and D. D. Feng, "Location classification of lung nodules with optimized graph construction," in *Proceedings of the 9th IEEE International Symposium on Biomedical Imaging (ISBI '12)*, pp. 1439–1442, Barcelona, Spain, May 2012.

[7] A. Farag, S. Elhabian, J. Graham et al., "Toward precise pulmonary nodule descriptors for nodule type classification," *Medical Image Computing and Computer-Assisted Intervention*, vol. 13, part 3, pp. 626–633, 2010.

[8] F. Zhang, Y. Song, W. Cai et al., "A ranking-based lung nodule image classification method using unlabeled image knowledge," in *Proceedings of the IEEE 11th International Symposium on Biomedical Imaging (ISBI '14)*, pp. 1356–1359, Beijing, China, May 2014.

[9] C. Jacobs, E. M. van Rikxoort, J.-M. Kuhnigk et al., "Automated characterization of pulmonary nodules in thoracic CT images using a segmentation-based classification system," in *European Congress of Radiology*, 2013.

[10] T. W. Way, B. Sahiner, H.-P. Chan et al., "Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features," *Medical Physics*, vol. 36, no. 7, pp. 3086–3098, 2009.

[11] R. Samala, W. Moreno, Y. You, and W. Qian, "A novel approach to nodule feature optimization on thin section thoracic CT," *Academic Radiology*, vol. 16, no. 4, pp. 418–427, 2009.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] F. Maldonado, J. M. Boland, S. Raghunath et al., "Noninvasive characterization of the histopathologic features of pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY)—a pilot study," *Journal of Thoracic Oncology*, vol. 8, no. 4, pp. 452–460, 2013.

[14] Y. Song, W. Cai, H. Huang, Y. Zhou, Y. Wang, and D. D. Feng, "Locality-constrained subcluster representation ensemble for lung image classification," *Medical Image Analysis*, vol. 22, no. 1, pp. 102–113, 2015.

[15] F. Zhang, W. Cai, Y. Song, M.-Z. Lee, S. Shan, and D. Dagan, "Overlapping node discovery for improving classification of lung nodules," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '13)*, pp. 5461–5464, Osaka, Japan, July 2013.

[16] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[17] ELCAP Public Lung Image Database, http://www.via.cornell.edu/databases/lungdb.html.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 2962047, 13 pages
http://dx.doi.org/10.1155/2016/2962047

Research Article

Automatic Approach for Lung Segmentation with Juxta-Pleural Nodules from Thoracic CT Based on Contour Tracing and Correction

Jinke Wang[1] and Haoyan Guo[2]

[1]*Department of Software Engineering, Harbin University of Science and Technology, Rongcheng, China*
[2]*School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China*

Correspondence should be addressed to Haoyan Guo; ghyghy@hit.edu.cn

Received 14 August 2016; Accepted 24 October 2016

Academic Editor: Ayman El-Baz

This paper presents a fully automatic framework for lung segmentation, in which juxta-pleural nodule problem is brought into strong focus. The proposed scheme consists of three phases: skin boundary detection, rough segmentation of lung contour, and pulmonary parenchyma refinement. Firstly, chest skin boundary is extracted through image aligning, morphology operation, and connective region analysis. Secondly, diagonal-based border tracing is implemented for lung contour segmentation, with maximum cost path algorithm used for separating the left and right lungs. Finally, by arc-based border smoothing and concave-based border correction, the refined pulmonary parenchyma is obtained. The proposed scheme is evaluated on 45 volumes of chest scans, with volume difference (VD) $11.15 \pm 69.63\,cm^3$, volume overlap error (VOE) $3.5057 \pm 1.3719\%$, average surface distance (ASD) 0.7917 ± 0.2741 mm, root mean square distance (RMSD) 1.6957 ± 0.6568 mm, maximum symmetric absolute surface distance (MSD) 21.3430 ± 8.1743 mm, and average time-cost 2 seconds per image. The preliminary results on accuracy and complexity prove that our scheme is a promising tool for lung segmentation with juxta-pleural nodules.

1. Introduction

Multidetector CT makes chest imaging with high-resolution and submillimeter isotropic characteristics, which greatly promote the automatic analytical techniques on medical imaging. Precise segmentation of pulmonary parenchyma is regarded as a critical step for automatic detection of various lung diseases. However, accurate lung segmentation often failed when abnormity turns up, and abnormity may be missed or other tissues that not belong to lungs could be included. Thus, conventional segmentation techniques are often insufficient to segment pulmonary parenchyma from chest CT datasets.

Previous work on lung segmentation can be roughly classified into two categories. The first category is threshold-based methods, which depend on the different attenuations between lung parenchyma and its surrounding tissues [1–9]. The main limitation of these methods is that their accuracy is badly influenced by pleura abnormity or artifact and often result in oversegmentation. Most of the threshold-based methods are two-dimensional approaches that process each axial section separately. Although it is a reasonable choice for thick slices CT, three-dimensional approach is more preferable when isotropic data is available, in which inconsistency between slices can be avoided.

Sun et al. [1] proposed a fully three-dimensional based lung segmentation and visualization technology. Firstly, in the preprocessing phase, isotropic filtering is used to improve the signal-to-noise ratio; and then, wavelet transform-based interpolation is applied to reconstruct the 3D voxels. Finally, by use of region growing, homogeneity, and gradient features, the lung region is extracted. Brown et al. [2, 3] also suggested a system framework based on 3D region growing and morphology smoothing; moreover, they proposed a semantic network anatomical model. On the basis of the attenuation threshold, shape, adjacent properties, volume, and relative position, the model can simulate the chest wall, mediastinum, bronchial tree, and left and right lungs to distinguish the different

anatomical structures. Sun et al. [4] developed a threshold-based segmentation method for missed diagnosis of large tumor. First, a normal shape model of lung is constructed by training of 41 sets of segmented datasets; second, for initialization, rib-based matching algorithm is used to produce the contour. Since the shape model cannot capture the details of the border, thus graph-cut method is implemented for the recovery of the details.

The second category is specific abnormity-based methods, which focus on specific abnormal diseases [10–15]. Due to their specificity on particular case, they are not applicable for routine test of large-scale datasets.

Sofka et al. [10] from Siemens used the visible structure knowledge of chest CT to present a multistage learning method. Firstly, the method identifies the spine among the tracheas; secondly, a hierarchical network is used to predict the posture parameter of left and right lungs. Thirdly, by use of the marks near the ribs and spine, a shape model is initialized and followed by a transformation operation to achieve the refinement. Korfiatis et al. [11] proposed a texture classification-based method for interstitial lung disease. The method used intensity-based K-means clustering for initialization, and for containing pixels that around the initial contour, the statistical features of intensity and wavelet coefficients are calculated for support vector classification. In order to compensate for the lost juxta-pleural nodules and ensure the smoothness of the lung boundary, several methods have been proposed to correct the lung contours [12, 13]. Yim and Hong [12] proposed a new curvature-based method for correcting the segmented lung boundary, a 3D branch-based region growing algorithm was utilized to segment the trachea and the left and right bronchi with adaptive growing conditions. Pu et al. [13] developed a lung segmentation method for reducing errors result from juxta-pleural tumor in traditional thresholding approaches. The proposed method begins with segmenting the lung contour with thresholding and smoothing and then flooding in the nonlung region of each slice; by this way, the initial border of the lung is tracked, and the adaptive border marching algorithm is utilized for reincluding the juxta-pleural tumor.

In addition to the above-mentioned studies, a few algorithms focus on diverse lung scans with dense pathologies being proposed. Sluimer et al. [16] proposed an atlas-based technology for lung segmentation with severe lesion. By registering 15 sets of chest CT to referenced lung atlas, the probability atlas is constructed, and then elastic registering is used for mapping the probability atlas to new scans for initialization and transformation. Finally, the trained lung border is utilized for refining the lung border.

The existing methods are either not taking the juxta-pleural tumors into consideration or too specified to be qualified for large-scale testing. Alleviating these difficulties is exactly what we are concerned with in this paper. We developed a fully automatic framework to segment pulmonary parenchyma with juxta-pleural nodules from chest CT. It starts from skin boundary detection with maximum connected component analysis, and then, rough segmentation of lung contour is implemented by diagonal-based tracing, which is followed by the separation of the left and right lungs with maximum cost path algorithm. And the final segmentation of pulmonary parenchyma is achieved by arc-based smoothing and concave-based correction. Our scheme is evaluated on 45 sets of CT scans, and its results are compared with the state of the art method, which is validated by the manual segmentation standard of radiologist.

2. Methods

In this section, the proposed framework will be described in detail. It is a multistep approach that gradually accumulates information until the final result is obtained. We depict the flowchart of the framework in Figure 1. It is subdivided into three phases: skin boundary detection, contour segmentation, and parenchyma refinement. In the rest of the parts, we further describe each individual step and explain how to segment pulmonary parenchyma automatically from chest CT.

2.1. Skin Boundary Detection. Skin boundary detection is the foundation of lung segmentation. In view of the high contrast between chest and the background, threshold-based method is utilized for segmentation purpose. In this section, firstly, principal component-based image aligning is implemented to correct the tilted scans; secondly, mathematical morphology operation is applied for noise reduction, and finally, by maximum connected region analyzing, the chest mask is extracted.

2.1.1. Principal Component-Based Image Aligning. The contour detection algorithm assumes that all patients have the same pose. In particular, it assumes that they lie upright and on their back in the scanner. This assumption is in most cases true due to the standardized CT scanning protocol. However, there are some rare cases in which the patients lie on their side, as shown in Figure 2(a). Because the border detection algorithm is not able to directly handle such scans in view of missing the starting point, an algorithm has been developed which automatically identifies scans in which patients lie on their side and rotates them accordingly.

In this paper, we limit the inclination angle on the x-y plane, and using the rotation method proposed by [17] for aligning. Firstly, for all the bone voxels on the x-y plane, principal component analysis [18] is applied for extracting the first principal component μ, and then μ is mapped to the positive direction of the x-axis to generate the rotation matrix R with the rotated degree ϕ:

$$\phi = \arctan\left(\frac{\mu_2}{\mu_1}\right), \tag{1}$$

where μ_1 is the mapping of vector μ on y-axis, while μ_2 is the mapping of vector μ on x-axis. It is assumed that μ is orthogonal to the patients sagittal plane and tangential to his coronal plane. It is further assumed that the angle ϕ between μ and the positive x-axis is between -90° and 90°. If this is not the case, that is, $\mu_1 < 0$, the direction of μ is inverted by multiplying -1.

A diagonal-based border detection algorithm is utilized in the subsequent section. By experience only if ϕ is out of

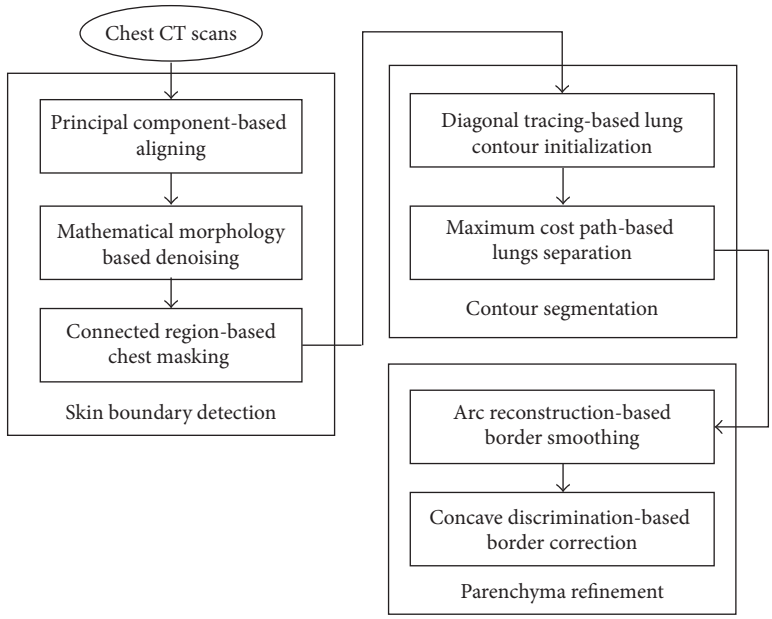
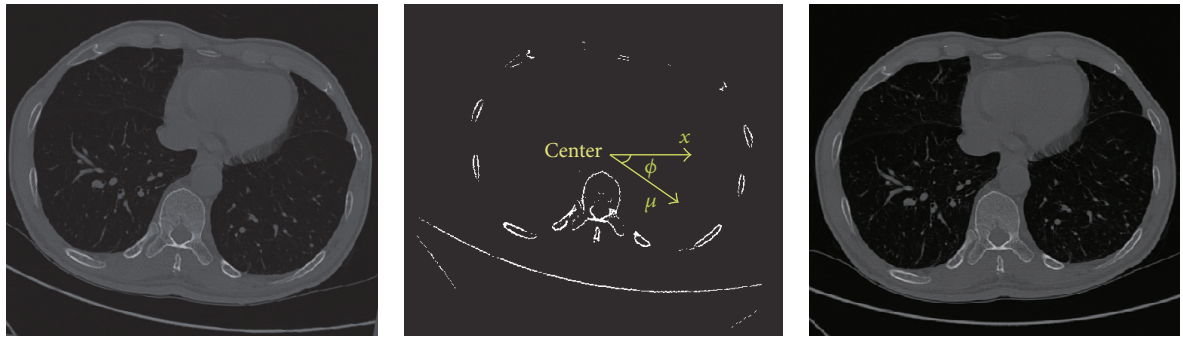


Figure 1: Flowchart of the proposed scheme on lung segmentation.



(a) (b) (c)

Figure 2: Illustration of image aligning. (a) Original tilted image. (b) Principal component analysis. (c) Aligned image after rotated.

$[-15, 15]$ can the aligning algorithm be applied, or the initial point of lung border could be missed. As ϕ is in $[-15, 15]$, the influence on the boundary tracking algorithm can be eliminated. As shown in Figure 2, by rotating around the center for ϕ degree, the tilted image is aligned.

2.1.2. Mathematical Morphology-Based Denoising. The main problem in skin boundary detection is the existence of various external noises, including human appendant, bed sheet, and CT scanner itself (Figure 3(a)). To eliminate these noises, firstly, Otsu threshold [19] is used for binary processing (Figure 3(b)); secondly, by morphological opening operation, salt noise in the CT scan, bed sheet, and the scanner itself are removed (Figure 3(c)). Finally, by connected regional analysis, the chest mask is determined (Figure 3(d)), and by masking the original chest scan, the final chest region is obtained (Figure 3(e)).

2.2. Rough Segmentation of Lung Contour. After skin boundary detection, we step into lung parenchyma segmentation. In this section, two procedures are applied: (1) diagonal tracing-based lung contour initialization; (2) maximum cost path-based lungs separation.

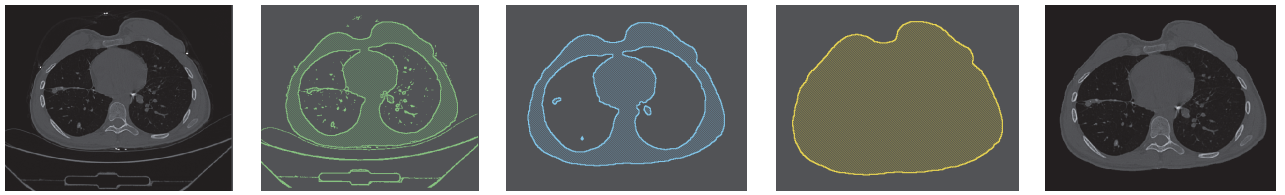
2.2.1. Diagonal Tracing-Based Lung Contour Initialization. A diagonal tracing-based method is proposed for lung contour initialization, with the detailed description in the following.

Step 1. Define the major diagonal as the searching path (Figure 4).

Step 2. Search first P_0 with three consecutive "0s" as the start point of the left lung.

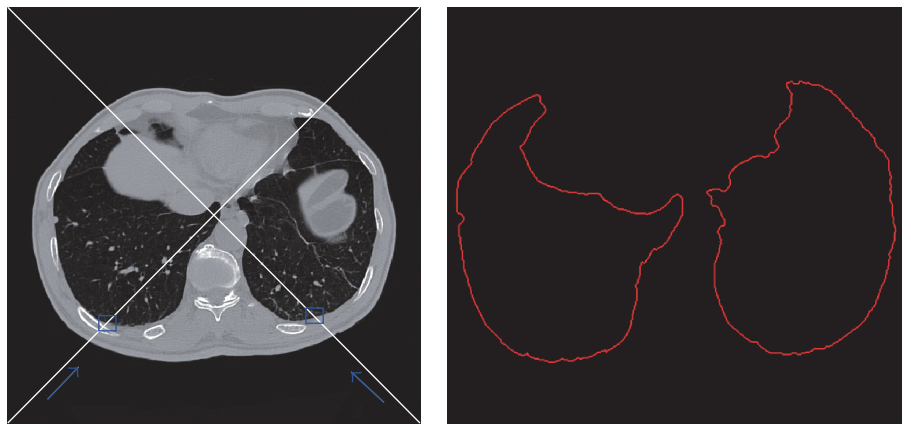
Step 3. 8-neighborhood-based boundary tracing is utilized for boundary extraction of the left lung. Assume the boundary point set is denoted by $\{P_1(a_1, b_1), P_2(a_2, b_2), \ldots, P_{n-1}(a_{n-1}, b_{n-1}), P_n(a_n, b_n)\}$.

Step 4. Once an overlap between the final two points and the initial two points is found, for example, $P_n = P_2$, $P_{n-1} = P_1$, the algorithm ends.



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

Figure 3: Illustration of skin boundary detection. (a) Original chest CT. (b) Otsu thresholding. (c) Morphological open. (d) Chest mask. (e) Final chest segmentation.



(a) (b)

Figure 4: Illustration of diagonal-based contour tracing. (a) Searching the start point along major and minor diagonal. (b) Rough contour after diagonal-based tracing.

By this way, the boundary of the left lung is achieved; similar to the method of obtaining the left lung border, by searching the start point along the minor diagonal, with the accompanied boundary tracing algorithm, the boundary of the right lung is achieved. Thus the initialization of lung contour is fulfilled.

2.2.2. Maximum Cost Path-Based Lungs Separation. The separation of left and right lungs is the necessary step for accurate lung segmentation. In [20, 21], 2D edge tracking was used to find the boundaries of the left and right lungs. Hu et al. [22] separated the left and right lungs by identifying the anterior and posterior junctions using dynamic programming. In this paper, we use the dynamic programming algorithm [22] for separation purpose. The dynamic programming algorithm is used on each slice with single connective component. Its target is to locate the position of the left and right lungs and reseparate them (see Figure 5). In this method, the weight map that is proportional to the intensity level is used for searching the maximum cost path, which corresponds to the separation line of left and right lungs.

Once the single connective area is found, 2D erosion process is applied for separation, while dilating process with constraint is used for reconstructing the original borderline. Supposing A as the original set of lung pixels, the erosion operation is adopted to calculate a new set for separated lungs S. The equation is showed as follows:

$$S = A \ominus nB_4, \tag{2}$$

where $\ominus$ indicates binary morphology erosion, and B_4 is a binary diamond-shaped structure. By iterative erosion with B_4, S is separated into two components, and the iterative number is indicated by n.

For the reconstruction of lung border, iterative dilation with constraint is used that is described as follows:

$$C^{i+1} = C^i \cup \{\{p\} \oplus B_4\}, \tag{3}$$

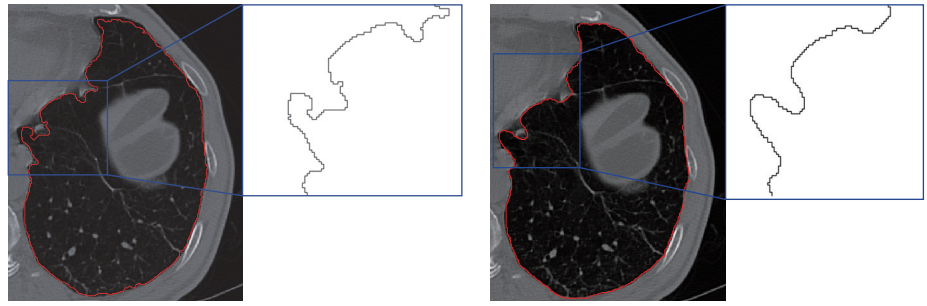
where $\oplus$ represents morphology dilation, with constraint $p \in C^i \cap A$, while C^i keeps the same components number with C^{i+1}, and $C^0 = S$ is used for initialization. Equation (3) is implemented until $p \in C^i \cap A$ is not satisfied or the component number is changed. Figure 5 illustrates the reconstruction process.

2.3. Pulmonary Parenchyma Refinement. In this section, two successive phases are implemented to refine the rough lung contour. We will describe the details step by step until the final pulmonary parenchyma is achieved.

2.3.1. Arc Reconstruction-Based Border Smoothing. Lots of jagged edges are generated after rough segmentation of lungs as shown in Figure 6. In order to make image smooth and reduce the impacts of gradient mutations, curve smoothing method is used. The partial arc coefficient is produced by multiple points, and through appropriate smoothing frequency, the optimum result is obtained. Since any curve on a

(a) (b)

Figure 5: Illustration of left and right lungs separation. (a) Connective case of left and right lungs. (b) Separation of left and right lungs after maximum cost path process.



(a) (b)

Figure 6: Processing of jagged border. (a) Image with jagged border. (b) Image after smoothing.

plane can be defined as $x = x(s)$, $y = y(s)$ (where s represents the arc length of the curve), therefore, the edge of the lung parenchyma can be denoted using (4):

$$\begin{aligned} x = x(s) &= a_0 + a_1 s + \cdots + a_n s^n, \\ y = y(s) &= b_0 + b_1 s + \cdots + b_n s^n, \end{aligned} \tag{4}$$

$$\begin{aligned} (P_n - P_1)^2 &\le 2, \\ (a_n - a_1)^2 + (b_n - b_1)^2 &\le 2. \end{aligned} \tag{5}$$

Smoothing is essentially a resampling process, and the convergence condition is (5). When the start points with the last two ones constitute a 8-neighborhood relation, a closed contour is determined. In this paper, cubic spline interpolation [23] is used for constructing the new smoothing border, and the detailed algorithm is described below.

Step 1. Resampling the initial contour ($\{P_1, P_2, \ldots, P_N\}$) with step size L, and then, the arc length between two adjacent points can be denoted by L; in this paper, $L = 0.3$ is used.

Step 2. For P_i on the border with adjacent points $\{P_{i-k}, \ldots, P_{i-1}, P_{i+1}, \ldots, P_{i+k}\}$ ($2k$ ($2k > n$)). The arc between the $2k + 1$ neighbors and P_i are $0, \ldots, (k-1)L, kL, (k+1)L, \ldots, 2kL$. Assuming (x_{i+j}, y_{i+j}) as the coordinate of P_{i+j}, and S_{i+j} as the arc length between P_{i+j} and P_{i-k}, we get the following deduction:

$$\begin{aligned} x_{i+j} &= a_0 + a_1 s_{i+j} + \cdots + a_n s_{i+j}^n, \\ y_{i+j} &= b_0 + b_1 s_{i+j} + \cdots + b_n s_{i+j}^n. \end{aligned} \tag{6}$$

Then, the least squares method [24] is utilized to obtain the coefficient series $a_0, a_1, \ldots, a_n, b_0, b_1, \ldots, b_n$.

Step 3. Take the arc lengths $s = kL$ of P_i and P_{i-k} into polynomial, and a new smoothed location $(\overline{x_i}, \overline{y_i})$ is generated.

Step 4. Repeat Steps 2 and 3 for a new border set until convergence.

Step 5. Set threshold T_2 for perimeter convergence, iterating from Steps 1 to 5 until $|C - C'| < T_2$.

The effect on jagged border smoothing is shown in Figure 6, with the testing parameters provided in Table 1. In this paper, parameters $n = 2$, $k = 4$, and $M = 12$ are selected.

2.3.2. Concave Discrimination-Based Border Correction. After border smoothing, appropriate detection and correction are required for solving undersegmentation problem caused by juxta-pleural nodules. The following approach is aiming to this target.

Concave area is defined as the line between the start point and the rightmost or leftmost point of step size. To

Table 1: Parameters in arc reconstruction-based smoothing.

	$n = 1$			$n = 2$			$n = 3$			$n = 4$	
k	M	t	k	M	t	k	M	t	k	M	t
1	50	14.2	2	120	23.3	2	120	30.5	3	240	56.6
2	10	5.30	3	30	6.81	3	30	6.57	4	135	43.4
3	8	15.4	4	12	3.25	4	10	2.79	5	50	17.6
4	4	0.78	5	6	1.71	5	6	1.57	6	20	9.10
5	2	0.62	6	4	1.09	6	4	1.15	7	15	5.71
	$n = 5$			$n = 6$			$n = 7$			$n = 8$	
k	M	t	k	M	t	k	M	t	k	M	t
3	300	98.06	4	620	252.33	4	800	380.52	5	1000	756.61
4	150	53.76	5	380	176.24	5	600	261.17	6	700	437.78
5	65	27.13	6	190	82.35	6	300	159.47	7	200	127.86
6	30	11.34	7	110	59.07	7	150	76.33	8	90	61.65
7	10	4.11	8	60	28.56	8	70	41.96	9	50	38.12

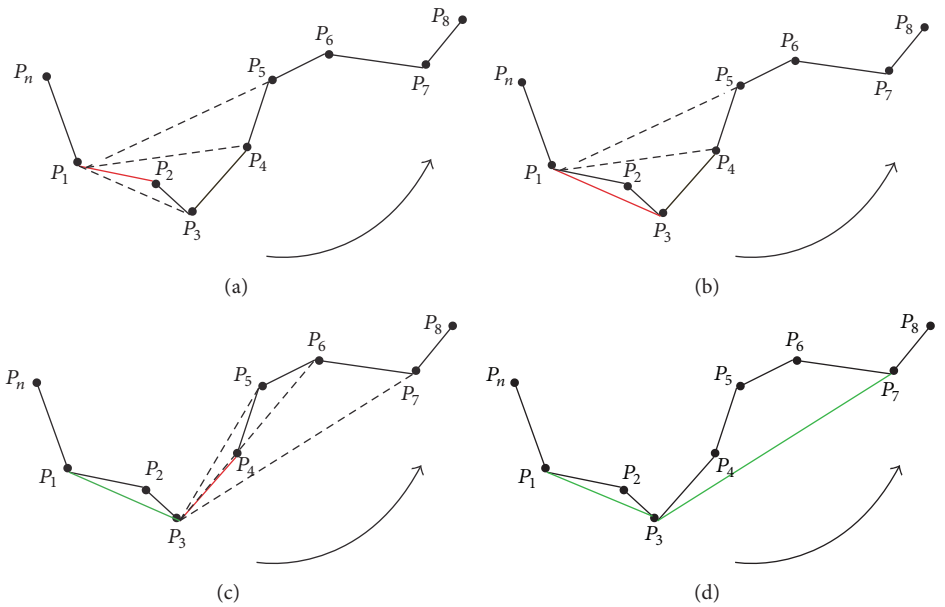


Figure 7: Illustration of border marching algorithm. (a) Start point. (b) Reference direction. (c) New point is found on the right of the reference direction. (d) New reference direction.

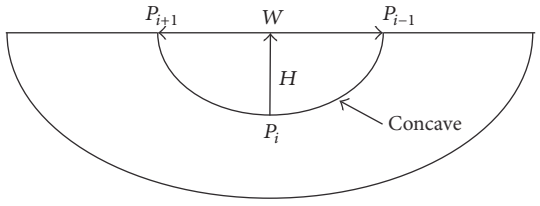
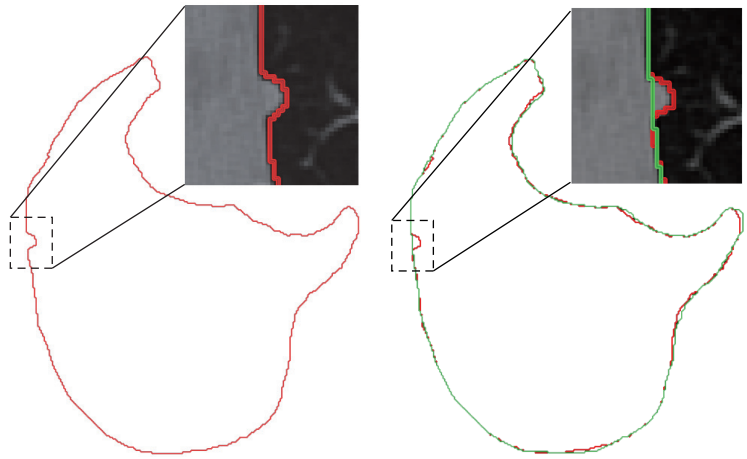


Figure 8: Illustration of the length and width of concave area, where W represents width, and H represents height.

determine the orientation, the right hand rule [25] is used, and for detecting all the possible concave areas, the adaptive border marching algorithm (ABM) [13] is utilized. We have developed a model with two parameters (Figure 8) for the determination of boundary refinement. One parameter is W, the Euclidean distance between two consecutive points after the marching operation, and the other is H, the maximum height perpendicular to this connecting line segment. We defined the threshold which is the length-width ratio of H and W. For any concave region where threshold $> T_1$, replace the concave area with a straight line. The ABM algorithm involves five consecutive points, as shown in Figure 7. Choose P_1 as the start point and P_1P_2 as the reference direction (indicated by red line); then, because point P_3 is found on the right of P_1P_2, thus, P_1P_2 is substituted by P_1P_3 as a new direction. Since all the rest points locate on the left side, thus, a new concave point is detected (indicated by green line). Then, a new round with the new point P_3 is continued until a closed path is achieved.

As concave region detection is completed, we step into the correction phase. On the one hand, concave area correction



(a) (b)

Figure 9: Illustration of border correction. (a) Undersegmentation. (b) After border correction Red line denotes the rough segmentation, while green line represents the effect of correction.

Table 2: Quality and accessibility of the image datasets.

Dataset	Number	Size	Resolution	Slices number	Slice thickness
Local hospital	45	512 ∗ 512	0.625–0.742 (mm)	275–502	0.55–1 (mm)

can reduce the missed diagnosis rate of juxta-pleural nodules; on the other hand, excessive correction will undoubtedly results in more undersegmentation errors. Therefore, length-width ratio-based threshold is proposed for solving this problem. The main procedure of this algorithm is described below.

Step 1. Calculate the perimeter C of lung border set P_1.

Step 2. For all the concave points on border set P_1, calculate the length-width ratio $\eta = H/W$ (see Figure 8).

Step 3. For any concave point that $\eta > T_1$, substitute the concave area with a straight line, where T_1 indicates the ratio-based threshold.

Step 4. Recalculate a new lung border set P_2 with perimeter C'.

Step 5. Set the threshold T_2 for perimeter convergence, iterating from Steps 1 to 5 until $|C - C'| < T_2$.

In this paper, the convergence threshold $T_2 = 0.01$ is used, and Figure 9 depicts the undersegmentation case via concave correction.

3. Experimental Results and Discussion

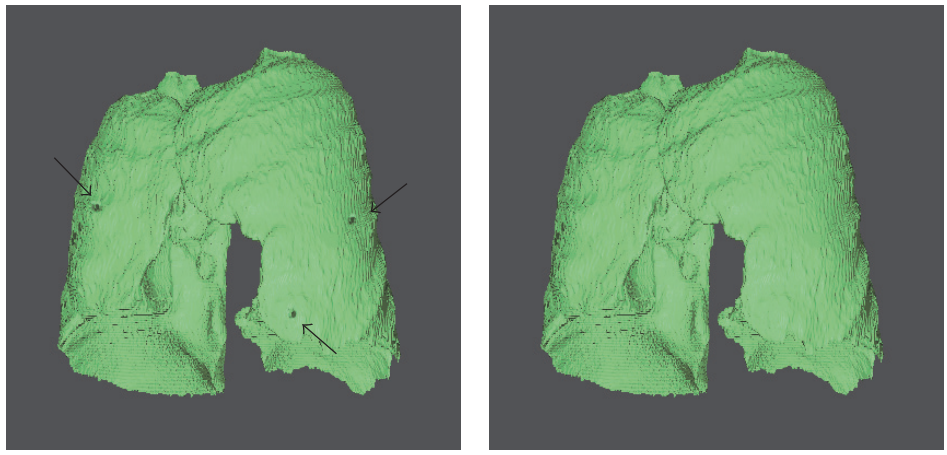
3.1. Quality and Accessibility of Datasets. A total of 45 sets of chest CT scans from Weihai Municipal Hospital are used for experiment, in which 20 groups are generated by Somatom Sensation 64 of Siemens Medical Systems, and another 25 groups come from Brilliance 64-bit scanner of Philips Medical Systems. CT size is $512 \times 512 \times 275$ to $512 \times 512 \times 502$, with pixel size 0.625 mm to 0.742 mm, and slice thickness 0.55 mm to 1.0 mm. Table 2 presents a detailed information about the quality of the images; meanwhile, Table 3 provides a detailed description of juxta-pleura tumor used in our experiment. All 45 groups are manually segmented by a radiologist as golden standard. Experiments are performed in Matlab R2010a, with quad-core CPU i5-4590 and 8 G memory.

Table 3: Information of juxta-pleura tumors.

Dataset	Types	Number	Size
Local hospital	Normal nodules; GGO	53	6–17 (mm)

The ground truth of the segmentation used in this paper was obtained by the radiologists of the cooperative hospital, utilizing a manual segmentation software named MITK [26], which provides an open source and a graphical user interface developed by the German Center for Cancer Research. The general procedure for ground truth segmentation is as follows. First, a smoothing operation is selected for reducing the noises in the images; second, a three-dimensional region growing method is used to obtain an initial segmentation result; finally, the rough segmentation results were further optimized by the radiologists on the cross-sectional, sagittal, or coronal slice until the final segmentation results were satisfactory.

3.2. Evaluation Metrics and Criteria. To evaluate the segmentation performance, seven error metrics are used in this paper, which are often utilized for evaluating on the



(a) (b)

Figure 10: 3D view of concave correction. (a) 3D view of juxta-pleural nodules before correction. (b) 3D view of juxta-pleural nodules after correction.

accuracy and complexity [6, 12], including volume difference (VD), volume overlap error (VOE), relative volume difference (RVD), average surface distance (ASD), root mean square distance (RMSD), maximum symmetric absolute surface distance (MSD), and process time. For automatic segmentation volume V_{auto} and manual segmentation volume V_{manu}, VD is defined as $\text{VD} = V_{\text{auto}} - V_{\text{manu}}$, $\text{RVD} = 100 \times ((V_{\text{auto}} - V_{\text{manu}})/V_{\text{ref}})$, $\text{VOE} = 100 \times (1 - (V_{\text{auto}} \cap V_{\text{ref}}/V_{\text{auto}} \cup V_{\text{manu}}))$, and ASD, RMSD, MSD are defined by (7), (8), and (9), respectively:

$$\text{ASD}(A,B) = \frac{1}{|S(A)| + |S(B)|}\left(\sum_{s_A \in S(A)} d(s_A, S(B)) + \sum_{s_B \in S(B)} d(s_B, S(A))\right), \tag{7}$$

$$\text{RMSD}(A,B) = \sqrt{\frac{1}{|S(A)| + |S(B)|}} \times \sqrt{\sum_{s_A \in S(A)} d^2(s_A, S(B)) + \sum_{s_B \in S(B)} d^2(s_B, S(A))}, \tag{8}$$

$$\text{MSD}(A,B) = \max\left\{\max_{s_A \in S(A)} d(s_A, S(B)), \max_{s_B \in S(B)} d(s_B, S(A))\right\}, \tag{9}$$

where A and B correspond to two segmentation results, and $d(v, S(X))$ represents the shortest Euler distance from voxel v to the segmentation result X.

Similar to the criteria used in [13], oversegmentation and undersegmentation rates are also considered as criteria for comparative study. Oversegmentation rate is defined as the segmentation volume that is regarded as lung tissue in our method, while not in the ground truth, and the undersegmentation rate is vice versa. We use the cumulative distribution to demonstrate the fitting between the lung surfaces obtained by the proposed method and the ground truth, which are calculated by the shortest distance between a point on the lung surfaces obtained by the proposed method and the lung surfaces of the ground truth.

3.3. Accuracy Analysis. Table 4 shows the experimental result on 45 chest scans, based on the proposed method and the golden standard. As shown in the table, VD is $11.15 \pm 69.63\,\text{cm}^3$, VOE is $3.5057 \pm 1.3719\%$, ASD is 0.7917 ± 0.2741 mm, RMSD is 1.6957 ± 0.6568 mm, and MSD is 21.3430 ± 8.1743 mm. In clinical practice, VOE of 5% is generally considered as the most acceptable error, and therefore, the proposed method is capable of providing clinical assist.

The automatic segmentation results were compared with manual segmentation result of the radiologist. Whether a juxta-pleural nodule was correctly included or not was determined by a radiologist to see whether there are obvious defects in the segmentation due to juxta-pleural nodules. Figure 10 shows the three-dimensional view before and after border correction. It can be seen that the juxta-pleural nodules are reincluded after correction operation. However, to a certain extent, oversegmentation error is inevitable due to the overcorrection; thus, most of VD in Figure 11 are positive, which appear above the x-axis, denoting oversegmentation. The over- and undersegmentation are 29 and 16 sets, respectively; in other words, the probability of oversegmentation is almost twice of undersegmentation.

In order to study the average distance of segmentation error, we depict the cumulative probability distribution based on under- and oversegmentation, which are showed in Figure 12. In general, oversegmentation is defined as the lung volume that is regarded as lung tissue in our segmentation method while not in the reference standard, and the undersegmentation is vice versa. In this paper, the metric of RVD (relative volume difference) is used for determining whether a segmentation belongs to oversegmentation or undersegmentation. If RVD gets a positive value, we regard

Table 4: Experimental result of lung segmentation on 45 testing scans. AVG indicates average result, and SD is short for standard deviation.

ID	VD (cm^3)	VOE (%)	RVD (%)	ASD (mm)	RMSD (mm)	MSD (mm)	V_{auto} (cm^3)	V_{manu} (cm^3)
1	−47.31	3.0242	−1.31	1.099	1.7593	18.9573	3671.56	3624.25
2	38.30	2.7479	1.09	0.7605	1.5615	18.122	3485.20	3523.50
3	71.74	3.6033	2.14	1.0078	1.7504	21.177	3275.97	3347.71
4	41.67	1.6723	1.14	0.6651	2.3539	19.6505	3615.24	3656.91
5	−54.48	3.2045	−1.47	1.123	2.0554	28.2471	3769.77	3715.29
6	31.33	3.6496	1.08	0.8762	1.7107	13.6632	2873.39	2904.72
7	−50.56	4.6419	−1.46	0.8933	1.4647	20.7681	3505.04	3454.47
8	−38.36	2.2887	−1.16	0.5502	1.5762	13.2609	3332.93	3294.57
9	75.49	5.2518	1.52	1.0722	1.5287	17.7907	4881.78	4957.27
10	−62.66	3.1748	−1.68	0.9567	1.8444	15.0682	3796.92	3734.26
11	35.49	2.5787	1.22	1.0584	1.6721	13.6448	2881.47	2916.96
12	43.62	1.6744	1.61	1.0107	1.806	19.3329	2658.32	2701.95
13	−49.49	3.1636	−1.45	1.0369	1.5537	10.0705	3459.81	3410.32
14	59.68	5.4911	1.99	0.7718	1.7302	19.3278	2945.88	3005.56
15	45.51	2.748	1.13	1.005	1.6431	19.9296	3967.29	4012.79
16	29.85	3.3606	0.77	0.9802	0.9411	11.3013	3825.41	3855.26
17	64.09	2.4039	1.83	0.4677	0.7013	10.2471	3439.35	3503.43
18	−57.97	4.3397	−1.91	1.1241	1.3929	22.9175	3099.42	3041.45
19	7.98	5.5238	0.23	1.1006	1.9721	15.0682	3531.61	3539.59
20	30.77	3.6862	0.78	1.0243	1.7763	20.1077	3938.13	3968.90
21	82.82	1.7774	2.20	0.3712	0.7476	18.913	3685.22	3768.04
22	78.42	6.1284	2.09	0.5164	0.9187	37.7341	3681.38	3759.81
23	39.89	5.0326	1.13	0.6109	1.8158	30.304	3491.12	3531.01
24	75.16	1.7624	2.24	0.4399	1.1044	17.6525	3283.02	3358.19
25	−50.81	4.5438	−1.31	0.6719	2.3576	32.5025	3918.14	3867.33
26	92.24	5.9488	2.38	1.0348	3.4893	38.3925	3776.06	3868.30
27	15.43	2.3037	0.53	0.3511	0.9273	16.7561	2879.30	2894.73
28	−46.44	2.5115	−1.34	0.5523	1.5216	16.6689	3508.69	3462.25
29	119.99	4.2139	2.69	0.6718	2.0628	29.6165	4342.92	4462.90
30	−29.24	2.8285	−0.60	0.8594	1.8682	24.9089	4887.26	4858.02
31	21.64	1.7654	0.57	0.5437	1.1489	16.5399	3799.97	3821.61
32	148.21	5.2781	4.87	1.0058	2.9211	25.7359	2892.25	3040.47
33	87.09	6.0183	3.16	1.0896	2.5128	21.8625	2668.69	2755.78
34	−168.48	4.5721	−3.91	1.0939	2.3051	21.9811	4472.12	4303.64
35	51.99	1.8947	1.73	0.3234	1.1074	26.6914	2950.36	3002.35
36	−97.87	5.4153	−2.53	1.1019	2.7363	33.2542	3970.86	3872.99
37	122.95	3.7016	2.48	0.8564	1.8731	22.6337	4835.66	4958.61
38	−137.55	5.2533	−4.15	1.1919	3.0683	47.3566	3449.08	3311.53
39	−83.38	2.6188	−2.75	0.6592	1.4036	17.7262	3110.78	3027.40
40	31.08	1.6871	0.87	0.4472	0.9098	13.2845	3543.87	3574.96
41	26.11	2.4515	0.66	0.2712	0.7285	26.0694	3945.08	3971.20
42	−71.33	2.7375	−1.97	0.7256	1.4599	17.6144	3692.75	3621.42
43	22.82	3.3868	0.62	0.4087	0.9147	10.2961	3681.11	3703.93
44	33.29	1.3839	0.94	0.3318	0.7743	12.3896	3505.11	3538.40
45	−77.18	4.3108	−2.40	0.9147	2.8345	34.8965	3290.29	3213.11
AVG	11.15	3.5057	0.3176	0.7917	1.6957	21.3430	3582.57	3593.71
SD	69.63	1.3719	1.9445	0.2741	0.6568	8.1743	529.47	528.37

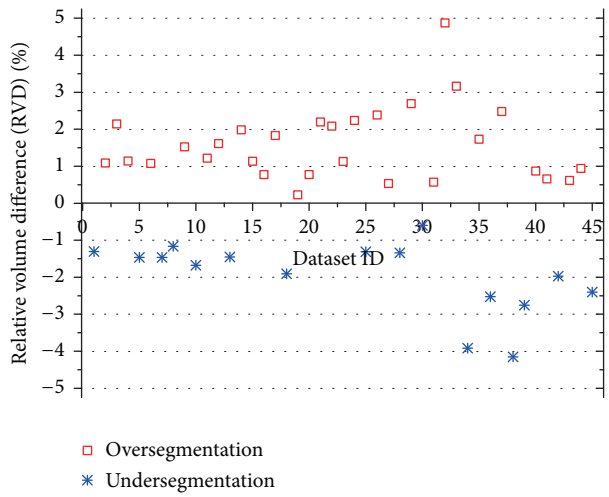


Figure 11: Comparative relative volume difference (RVD) of under-segmentation and oversegmentation.

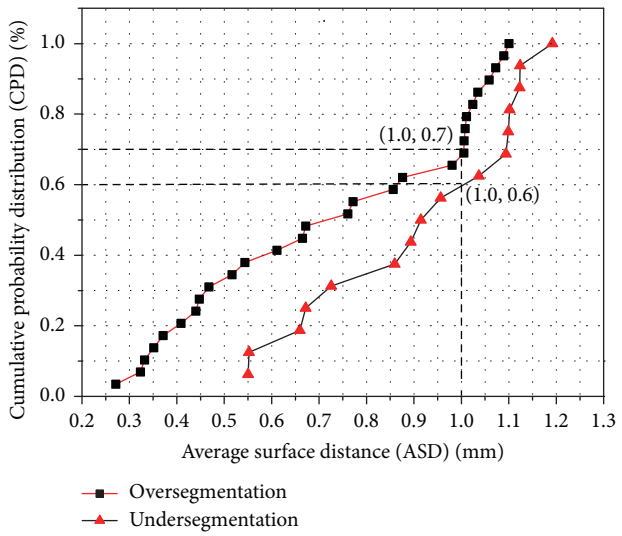


Figure 12: Cumulative probability distribution function of under-segmentation and over-segmentation. (■) 29 groups of over-segmentation. (▲) 16 groups of under-segmentation.

this segmentation result as an oversegmentation on the whole, or undersegmentation vice versa.

We use the cumulative distribution to demonstrate the fitting between the lung surfaces obtained by our method and the manual segmentation standard. The cumulative distance distribution is formed by calculating the metric of ASD (average surface distance) obtained by our automatic method and manual segmentation standard.

In Figure 12, cumulative probability distribution for over- and undersegmentation within 1 mm are 70% and 60%, respectively, while the maximum distance errors are 1.1 mm and 1.2 mm, respectively, thus proving the higher probability of segmentation errors generated by oversegmentation.

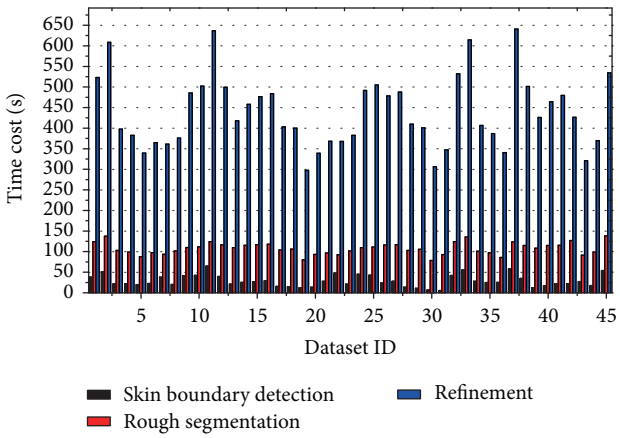


Figure 13: Time-consuming diagram of the main phases of the whole system.

3.4. Complexity Analysis. Figure 13 shows the time-consuming diagram of the main processing phases, including skin boundary detection, rough segmentation lung parenchyma, and refinement of lung parenchyma. In the figure, the whole time-consuming is 537.73 ± 162.873 seconds on average, among which skin boundary detection costs 25.2 ± 12.2376 seconds (accounting for 4.66% of the whole time), rough segmentation of lung parenchyma costs 102.45 ± 28.5473 seconds (accounting for 19.03% of the whole time), and refinement costs 408.98 ± 16.788 seconds (accounting for 76.31% of the overall time). And the time complexity for these phases are $\mathcal{O}(N^2)$, $\mathcal{O}(k_1N^2)$, and $\mathcal{O}(k_2N^2)$, respectively, where k_1 indicates the iteration numbers of maximum cost path, while k_2 indicates the iterative number of reconstruction and concave determination.

It can be seen from Figure 13 that the proposed scheme spends much more time on smoothing and correction process, in which iterative convergence accounts for a large proportion. On the average, processing time for each image is 2 seconds, while radiologist needs 1 minute for manual segmentation, which proves the efficiency of the proposed scheme.

3.5. Comparison with State of the Art Method. To evaluate the performance of our method, the proposed method was compared with the state of the art method proposed by Pu et al. [13]. In Pu's method, an adaptive border marching (ABM) algorithm was proposed to segment the lung and correct the segmentation defects caused by juxta-pleura nodules while minimizing undersegmentation and oversegmentation relative to the true lung border. The primary emphasis and distinguishing characteristic of the proposed method is on robustly correcting missed juxta-pleural nodules.

Table 5 presents the lung segmentation results by using our proposed method on 45 datasets, when compared to the conventional ABM-based method (Pu's method). For the final segmentation results, our method yields mean VOE of 3.51% and ASD of 0.79 mm, while conventional ABM-based method yields mean VOE of 3.86% and ASD of 0.83 mm. Our

Table 5: Comparative results between our method and the conventional method.

Method	VOE [%]	ASD [mm]	RMSD [mm]	MSD [mm]
Conventional ABM	3.8574 ± 2.10	0.8304 ± 0.33	1.8783 ± 0.51	32.2461 ± 8.12
Our method	3.5057 ± 1.94	0.7917 ± 0.27	1.6957 ± 0.66	21.3430 ± 8.17

Table 6: Relative volume difference (RVD) results between our method and the conventional method.

Method	RVD (oversegmentation) [%]	RVD (undersegmentation) [%]
Conventional ABM	1.92 ± 1.02	−1.65 ± 0.93
Our method	1.87 ± 0.95	−1.58 ± 0.96

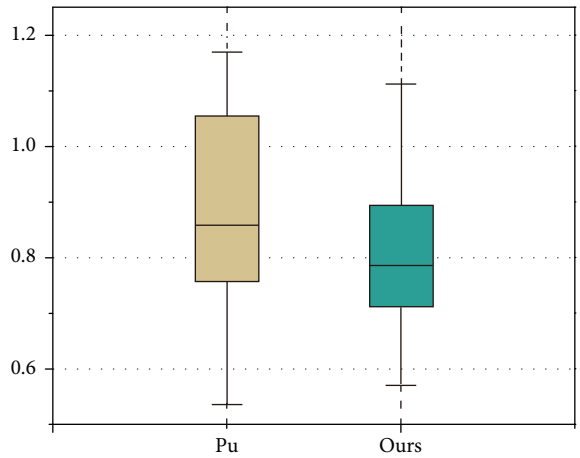


Figure 14: Boxplot on ASD between Pu's and our methods.

method outperforms the conventional method by 0.35% and 0.04 mm on average in terms of AOE and ASD, respectively.

In addition, similar results were also obtained for comparison study by boxplot between Pu's and our method in Figure 14. For the ASD, there is no abnormal point in both results; meanwhile, p value is less than 0.05 by t test, hence proving the significant better accuracy of our method.

Table 6 lists the comparative RVD results on both methods. For the RVD result of oversegmentation, our method and conventional ABM yield mean 1.87% and 1.92%, respectively, while, for the RVD result of undersegmentation, our method and conventional ABM yield mean −1.58% and −1.65%, respectively. Our method outperforms the conventional ABM by 0.05% and 0.07% on average in terms of RVD on oversegmentation and undersegmentation, respectively.

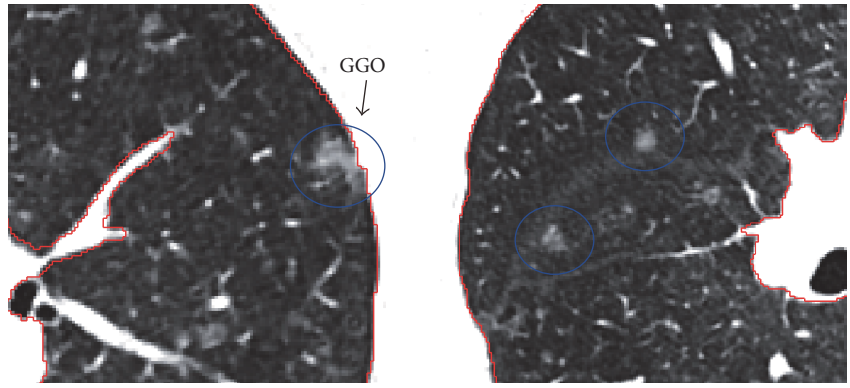
Therefore, for the lung tissue with juxta-pleura nodules, our proposed method achieves more accurate and robust segmentation results than the conventional method. The main reason for that is the lungs separation operation in our method improves the accuracy of lung contour segmentation. It can thus be deployed for accurate and robust lung segmentation with juxta-pleura nodules.

For our study, the main target is to solve the problem of the segmentation result by juxta-pleural nodules; thus it is not a generic tool to have this segmentation method when lung includes other pathological lesions or abnormities especially near the pleura. However, in our datasets, GGO (ground glass opacity) nodules are also considered in our scans, some of which are attached to the pleura. Although GGO often shows the irregular shape and low intensity, and its irregular shape is usually not fit to regular concave area detection algorithm, its low contrast with the lung parenchyma helps obtain the correct lung contour. In Figure 15(a) the lung segmentation is performed correctly due to the superiority of border tracing algorithm even when GGO is attached to the pleura, while other conventional region growing-based methods often need further processing because of the inhomogeneity between GGO and the lung parenchyma.

We recognize that the proposed scheme still needs further improvement. As the failure cases Figure 16(a) demonstrated, oversegmentation occurred around the trachea which is close to the parenchyma, and that is the result of overcorrection. In Figure 16(b), when the big vessel is located on the edge of the lung parenchyma, undersegmentation occurred because of the undercorrection. It is difficult to overcome this dilemma through 2D slice since border correction is a trade-off problem. Nevertheless, the segmentation error generated by juxta-pleura nodules could be reduced significantly due to the appropriate length-to-high ratio. Further study on how to reduce the errors caused by trachea and vessel could help alleviate the above-mentioned dilemma.

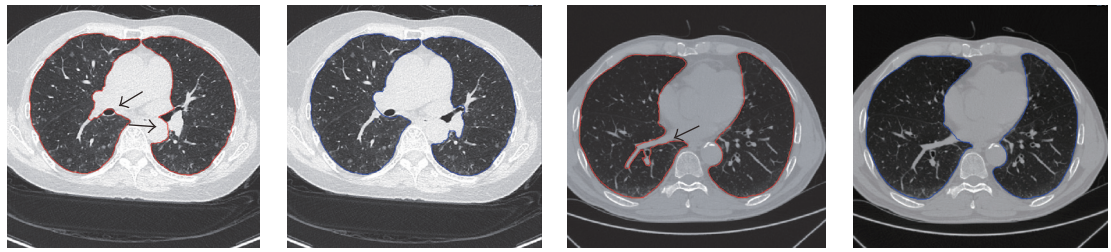
4. Conclusion

In this paper, a fully lung segmentation framework for chest CT with juxta-plural nodules is proposed via five main procedures, including chest segmentation, lung border tracing, left and right lung separation, lung border smoothing, and border correction, which focus on the oversegmentation problem caused by juxta-pleural nodules. Compared with manual segmentation, the volume overlap error of our approach is less than 5%, which meets the clinical requirements. And also, the time-consuming is about 2 seconds per image, which is more efficient than the manual cost of 1 minute per image. However, the proposed scheme tends to result in some undersegmentation, especially around the area which is close to the mediastinum, where the dense tracheas are located. Therefore, the border correction algorithm still needs further improvement, especially for the irregular lung shape caused by abnormal lesions. Nevertheless, comparing with the traditional method, our proposed scheme achieved great



(a) (b)

Figure 15: Illustration of lung segmentation result with GGO (ground glass opacity) nodules. (a) Juxta-pleura GGO. (b) Isolated GGO.



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

Figure 16: Illustration of segmentation errors of undersegmentation and oversegmentation. (a) and (c) indicate oversegmentation and undersegmentation, respectively. (b) and (d) indicate the manual segmentation as golden standard.

advantages in accuracy and time complexity, which indicates a potential tool for lung segmentation with juxta-pleural nodules.

Consent

Informed consent was obtained from all patients for being included in the study.

Competing Interests

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

Acknowledgments

This work was supported by the Nature Science Foundation of Heilongjiang Province of China (no. QC2016090) and the Science and Technology Planning Fund of Colleges and Universities of Shandong Province of China (no. J16LN60).

References

[1] X. Sun, H. Zhang, and H. Duan, "3D computerized segmentation of lung volume with computed tomography," *Academic Radiology*, vol. 13, no. 6, pp. 670–677, 2006.

[2] M. S. Brown, M. F. McNitt-Gray, N. J. Mankovich et al., "Method for segmenting chest CT image data using an anatomical model: Preliminary results," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 828–839, 1997.

[3] M. S. Brown, J. G. Goldin, M. F. McNitt-Gray et al., "Knowledge-based segmentation of thoracic computed tomography images for assessment of split lung function," *Medical Physics*, vol. 27, no. 3, pp. 592–598, 2000.

[4] S. Sun, C. Bauer, and R. Beichel, "Automated 3-D segmentation of lungs with lung cancer in CT data using a novel robust active shape model approach," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 449–460, 2012.

[5] S. Ukil and J. M. Reinhardt, "Smoothing lung segmentation surfaces in three-dimensional X-ray CT images using anatomic guidance," *Academic Radiology*, vol. 12, no. 12, pp. 1502–1511, 2005.

[6] E. M. Van Rikxoort, B. De Hoop, M. A. Viergever, M. Prokop, and B. Van Ginneken, "Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection," *Medical Physics*, vol. 36, no. 7, pp. 2934–2947, 2009.

[7] A. Hasegawa, S.-C. B. Lo, J.-S. Lin, M. T. Freedman, and S. K. Mun, "A shift-invariant neural network for the lung field segmentation in chest radiography," *Journal of Signal Processing Systems*, vol. 18, no. 3, pp. 241–250, 1998.

[8] J. K. Leader, B. Zheng, R. M. Rogers et al., "Automated lung segmentation in X-Ray computed tomography: development and evaluation of a heuristic threshold-based scheme," *Academic Radiology*, vol. 10, no. 11, pp. 1224–1236, 2003.

[9] Y. Yang, S. Zhou, P. Shang, E. Qi, S. Wu, and Y. Xie, "Contour propagation using feature-based deformable registration for

lung cancer," *BioMed Research International*, vol. 2013, Article ID 701514, 8 pages, 2013.

[10] M. Sofka, J. Wetzl, N. Birkbeck et al., "Multi-stage learning for robust lung segmentation in challenging ct volumes," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2011: 14th International Conference, Toronto, Canada, September 18–22, 2011, Proceedings, Part III*, vol. 6893 of *Lecture Notes in Computer Science*, pp. 667–674, Springer, Berlin, Germany, 2011.

[11] P. Korfiatis, C. Kalogeropoulou, A. Karahaliou, A. Kazantzi, S. Skiadopoulos, and L. Costaridou, "Texture classification-based segmentation of lung affected by interstitial pneumonia in high-resolution CT," *Medical Physics*, vol. 35, no. 12, pp. 5290–5302, 2008.

[12] Y. Yim and H. Hong, "Correction of segmented lung boundary for inclusion of pleural nodules and pulmonary vessels in chest CT images," *Computers in Biology and Medicine*, vol. 38, no. 8, pp. 845–857, 2008.

[13] J. Pu, J. Roos, C. A. Yi, S. Napel, G. D. Rubin, and D. S. Paik, "Adaptive border marching algorithm: automatic lung segmentation on chest CT images," *Computerized Medical Imaging and Graphics*, vol. 32, no. 6, pp. 452–462, 2008.

[14] S. Zhou, Y. Cheng, and S. Tamura, "Automated lung segmentation and smoothing techniques for inclusion of juxtapleural nodules and pulmonary vessels on chest CT images," *Biomedical Signal Processing and Control*, vol. 13, no. 1, pp. 62–70, 2014.

[15] M. N. Prasad, M. S. Brown, S. Ahmad et al., "Automatic segmentation of lung parenchyma in the presence of diseases based on curvature of ribs," *Academic Radiology*, vol. 15, no. 9, pp. 1173–1180, 2008.

[16] I. Sluimer, M. Prokop, and B. Van Ginneken, "Toward automated segmentation of the pathological lung in CT," *IEEE Transactions on Medical Imaging*, vol. 24, no. 8, pp. 1025–1038, 2005.

[17] M. Kirschner, *The probabilistic active shape model: from model construction to flexible medical image segmentation [Ph.D. thesis]*, TU Darmstadt, 2013.

[18] I. T. Jolliffe, *Principal Component Analysis*, Wiley Online Library, 2002.

[19] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285–296, pp. 23–27, 1975.

[20] W. A. Kalender, H. Fichte, W. Bautz, and M. Skalej, "Semiautomatic evaluation procedures for quantitative ct of the lung," *Journal of Computer Assisted Tomography*, vol. 15, no. 2, pp. 248–255, 1991.

[21] L. W. Hedlund, R. F. Anderson, P. L. Goulding, J. W. Beck, E. L. Effmann, and C. E. Putman, "Two methods for isolating the lung area of a CT scan for density information," *Radiology*, vol. 144, no. 2, pp. 353–357, 1982.

[22] S. Hu, E. A. Hoffman, and J. M. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images," *IEEE Transactions on Medical Imaging*, vol. 20, no. 6, pp. 490–498, 2001.

[23] H. Späth, *Spline Algorithms for Curves and Surfaces*, Utilitas Mathematica Publishing, 1974.

[24] E. Hinton and J. S. Campbell, "Local and global smoothing of discontinuous finite element functions using a least squares method," *International Journal for Numerical Methods in Engineering*, vol. 8, no. 3, pp. 461–480, 1974.

[25] P. Bose, P. Morin, I. Stojmenović, and J. Urrutia, "Routing with guaranteed delivery in ad hoc wireless networks," *Wireless Networks*, vol. 7, no. 6, pp. 609–616, 2001.

[26] I. Wolf, M. Vetter, I. Wegner et al., "The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK," in *Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display, 16*, vol. 5367 of *Proceedings of SPIE*, May 2004.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 1067509, 10 pages
http://dx.doi.org/10.1155/2016/1067509

Research Article

CRF-Based Model for Instrument Detection and Pose Estimation in Retinal Microsurgery

Mohamed Alsheakhali,[1] Abouzar Eslami,[2] Hessam Roodaki,[1] and Nassir Navab[1]

[1] *Technische Universität München, Munich, Germany*
[2] *Carl Zeiss Meditec AG, Munich, Germany*

Correspondence should be addressed to Mohamed Alsheakhali; alsheakh@in.tum.de

Received 31 July 2016; Revised 25 September 2016; Accepted 3 October 2016

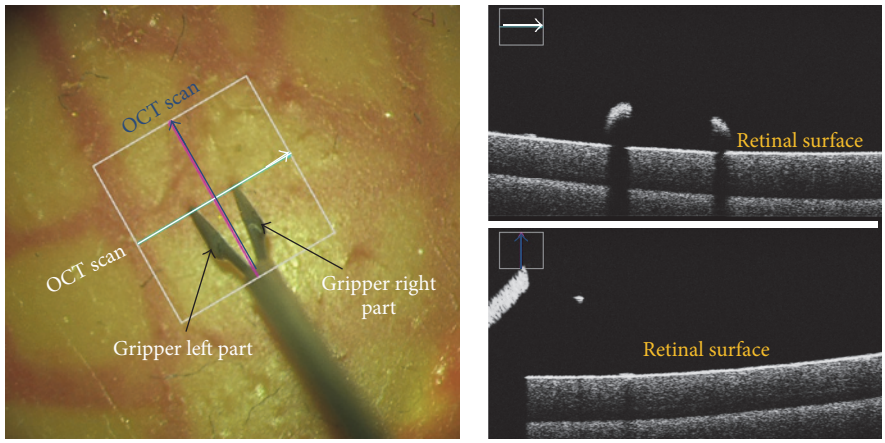
Academic Editor: Georgy Gimel'farb

Detection of instrument tip in retinal microsurgery videos is extremely challenging due to rapid motion, illumination changes, the cluttered background, and the deformable shape of the instrument. For the same reason, frequent failures in tracking add the overhead of reinitialization of the tracking. In this work, a new method is proposed to localize not only the instrument center point but also its tips and orientation without the need of manual reinitialization. Our approach models the instrument as a Conditional Random Field (CRF) where each part of the instrument is detected separately. The relations between these parts are modeled to capture the translation, rotation, and the scale changes of the instrument. The tracking is done via separate detection of instrument parts and evaluation of confidence via the modeled dependence functions. In case of low confidence feedback an automatic recovery process is performed. The algorithm is evaluated on in vivo ophthalmic surgery datasets and its performance is comparable to the state-of-the-art methods with the advantage that no manual reinitialization is needed.

1. Introduction

Retinal microsurgery is among the most delicate operations requiring microprecision medical instruments. Usage of such instruments is manually carried out by surgeons to manipulate retinal tissue. An efficient feedback of the distance between the instrument tip and the retina would minimize tissue damage caused by unintentional touch of retina. Recently, ophthalmic surgical microscopes are equipped with intraoperative Optical Coherence Tomography (OCT), which has been used in [1] to estimate the distance of the instrument to the retinal surface. However, continuous real-time detection of the instrument tip and instrument orientation is still required to enable automatic repositioning of the OCT scans during live surgery. Figure 1 depicts two OCT scans acquired at the instrument tip and in the orientation of its shaft. The proper position is marked with a cross. The first scan (white color) is positioned at the two tips, and the corresponding OCT cross-section is shown in the upper part of Figure 1(b) which shows the retinal surface and how far the two tips are from it. The second orthogonal scan (blue color) is positioned along the instrument shaft and the cross-section corresponding to this scan is shown in the lower part of Figure 1(b). The depth information here shows the retinal surface and depth of the forceps center point. Augmenting the scene with depth information in addition to the 2D coordinates of the instruments tips brings new advantages for minimally invasive procedures. Detecting and tracking the instrument tips are needed to provide the OCT device with the new position information, and it is the most challenging step, especially for forceps instruments. Many factors such as the cluttered background, presence of blood vessels, instrument shadow, and rapid illumination have negative impact on tracking quality. Recent approaches [2, 3] modeled the instrument as a multiparts object where the parts are connected to each other in a linear way. Such approaches do not have the ability to detect the instrument tips in case of forceps usage where the linearity condition of the parts distribution is not satisfied.

In this paper, a new instrument detection, tracking, and pose estimation solution is presented. This solution relaxes the linear configuration of the instrument's parts



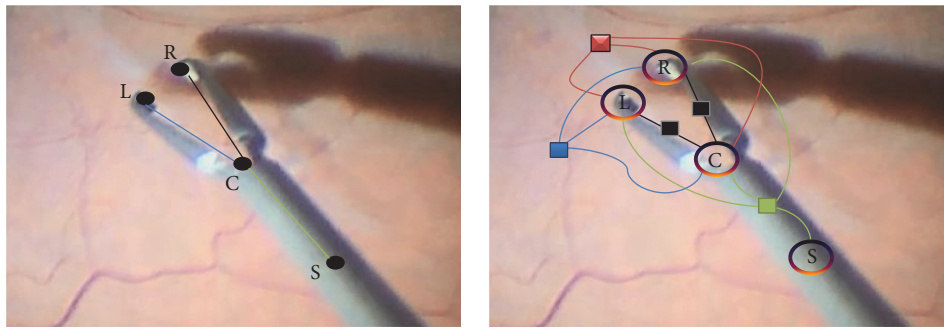
(a) (b)

Figure 1: (a) Microscopic image with two OCT scans in a cross sign; (b) OCT depth information along each scan.

and provides more robust model to handle different types of forceps used in eye surgery. Our method models the instrument as a Conditional Random Field (CRF) in which different parts of the instrument are detected in the 2D space of the image. Multiple models are implemented to capture the translation, rotation, and scale changes among the parts. One great advantage of the approach over the state-of-the-art methods is the ability to handle tracking failures in real-time. Such circumstances occur often in real complex datasets. The algorithm maintains confidence values to know whether to keep tracking by detection or to reinitialize the detection automatically. A second achievement of our approach is that it is the first proposed method, to the best of our knowledge, that can locate not only the instrument tips, but also its orientation in case of forceps instrument. Therefore, it provides all parameters needed to position OCT scans to get the distance between the tips and the retinal surface. Experimental results demonstrate the efficiency, robustness, and accuracy of our method in real in vivo scenarios and its ability to work on long videos. Comparisons with the state of the art on public and laparoscopic datasets demonstrate comparable results with the advantage that no manual reinitialization is needed.

2. Previous Work

Much research has been done to address the problem of detecting and tracking medical instruments including color-based [4, 5] and geometry-based [6–8] approaches. A recent work of Roodaki et al. [1] proposed to estimate the instrument tip depth to retina surface by building their method on top of instrument tracking algorithms. Despite of the high accuracy of the estimated depth, the algorithm relies on manual positioning of OCT scans or tracking algorithms which are prone to fail under high appearance changes. Many algorithms for instrument tracking and detection have been developed to be integrated with OCT depth estimation algorithms. However, there are many limitations of these algorithms preventing them to be used in real in vivo surgery. Sznitman et al. [9] proposed a unified framework to solve detection and tracking as a density estimation problem. The basis for this method is to model the instrument localization as a sequential entropy minimization problem to estimate 3DOF parameters required to localize the instrument tip. The method was evaluated using simple vitrectomy instrument, and it is not working on forceps used in retinal peeling operations. Therefore, such a method cannot localize the forceps two tips for automatic positioning of OCT scans. Modeling the instrument as multiple linearly connected parts was proposed in [2], but the linearity constraint limits its capabilities to detect only the center point in case of forceps instrument which is not sufficient for minimally invasive procedures. Machine learning based detectors [10] and online learning methods [11] have been employed to track only the center point of the instrument without detecting forceps tips. Reiter et al. [6] proposed a solution to track the instrument by making use of the landmarks on its surface. Color, location, and gradient-based features have been associated with the landmarks for training random ferns. The 3D locations of the instrument are retrieved by matching the features tracks in the stereo camera using normalized cross correlation. The method achieves high localization accuracy. However, it cannot run at the video frame rate due to the computational cost of extracting all these features. Moreover, the occlusions of some landmarks due to the instrument rotation might result in high localization error. Another approach [12] was proposed for articulated instrument tracking in 3D laparoscopic images, in which the color information is used for instrument parts segmentation. The segmented regions are described by different statistical models in order to estimate the pose of the instrument in the 3D space. Optical flow is used for pose tracking from image to another. The approach has also the limitations of expensive feature extraction and high sensitivity to the light changes. Rieke et al. [13] proposed to use regression forests to localize the forceps tips within a bounding box. However, this bounding box is provided using intensity-based tracker. Hence, once the tracker gets lost, the operation has to be interrupted to reinitialize OCT device manually. A recent work [3] proposed to use the deep learning



(a) (b)

Figure 2: (a) Target pose estimation; (b) the factor graph for the forceps: 4 variables (left (L), right (R), center (C), and shaft (S)) are used with different types of constraints that are presented with different edge colors: black (translation), green (rotation), red (relative length), and blue (consistency).

to detect the instrument parts and estimate its orientation. The approach achieved comparable results to the state-of-the-art methods but it is computationally expensive as well as it cannot detect the two forceps tips. Generally, most of the limitations are due to the time complexity or inability to detect forceps two tips and forceps orientation which are addressed in the work of this paper.

3. Proposed Method

Medical instrument, in this work, is modeled as multiparts articulated object where each part can be detected separately. Depending on the used features, parts detections using most of machine learning classifiers can result in a large number of false detections especially for structure-less objects like our target. However, these detections, including the true positive ones, form a new and reduced search space within the 2D image space which represents instrument part's hypotheses space. Therefore, the sought targets are just specific instrument part detections within the reduced space, such that the detected parts would represent the final instrument configuration. Prior information about the instrument parts and the relations between them are integrated on top of these detections together in one model in order to filter out the vast majority of false detections and to end up with the optimal instrument configuration. Prior instrument information can include the relative lengths of the parts, the angles between them, the gripper length, the possible movements of the joint, the possible changes of the current state, and so forth. Given different models, expressed as probabilistic distributions, to describe prior information about the instrument, and some potential instrument configurations, then the ultimate goal of our approach is to optimize for the best configuration (instrument pose) as shown in Figure 2(a) which maximizes the likelihood of the distributions of the prior models. To this end, the instrument in our method is modeled as a CRF of n random variables, and the factor graph of this model is shown in Figure 2(b). Each random variable Y_i corresponds to an instrument part, and the edges among these variables denote conditional dependence of the parts which can be described as a physical constraint. The instrument pose is given by the configuration $Y = (Y_1, Y_2, \ldots, Y_n)$, where the state for each variable $Y_i \in \Lambda_i$ represents the 2D position of the instrument part and is taken from the discrete space $\Lambda_i \subset R^2$. Consider an instance of the observation $x \in X$ that corresponds to instrument parts features, a reference pose P, and an instrument configuration $y \in Y$; the posterior is defined as

$$\begin{aligned} p(y \mid x, P) = \frac{1}{Z(x,P)} \prod_i^n \Phi_i^{\text{Conf}}(y_i, x) \cdot \Phi_i^{\text{Temp}}(y_i, P_i) \\ \cdot \prod_{(i,j)\in E_{\text{Trans}}} \Psi^{\text{Conn}}(y_i, y_j) \\ \cdot \prod_{(i,j,k)\in E_{\text{RLen}}} \Psi^{\text{RLen}}(y_i, y_j, y_k) \\ \cdot \prod_{(i,j,k)\in E_{\text{Cons}}} \Psi^{\text{Cons}}(y_i, y_j, y_k) \\ \cdot \prod_{(i,j,k,l)\in E_{\text{Rot}}} \Psi^{\text{Rot}}(y_i, y_j, y_k, y_l), \end{aligned} \tag{1}$$

where $Z(x, P)$ is the partition function and $\Phi^{\text{Conf}}(y_i, x)$ is the unary score function. E_{Trans}, E_{RLen}, E_{Cons}, and E_{Rot} are the graph edges that model the kinematic constraints among the instrument parts using different potentials functions. Ψ^{Conn} is binary potentials functions to model the distances changes among the forceps gripper's end points based on the connectivity between the forceps center point and each of the tips. Ψ^{RLen} and Ψ^{Cons} are ternary potentials functions to ensure consistency in the relative length of the left and right parts of the gripper and whether they can be bounded by a small region in the image. The rotation potential function Ψ^{Rot} is defined to estimate the configuration likelihood based on the distribution describing the proper angles among the instrument parts. Once the forceps hypothetical parts are detected, different configurations from these hypotheses within a defined Region of Interest (ROI) are evaluated with the potential functions to select one configuration. This configuration is the one maximizing the posterior given in (1) and it represents the forceps pose.

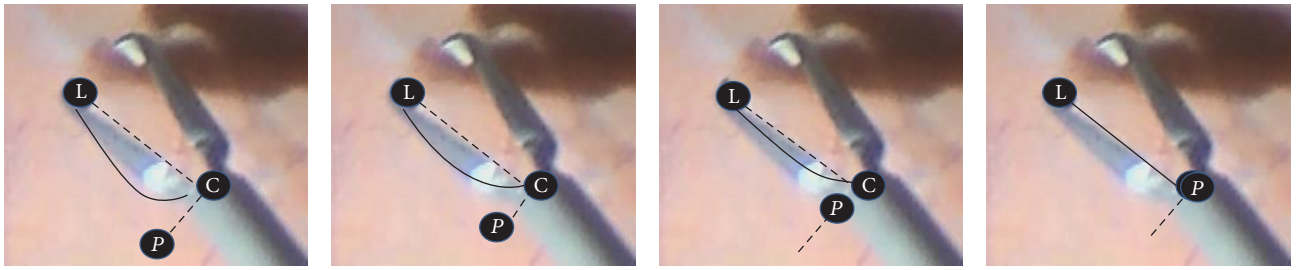


Figure 3: Connectivity modeling using Bézier curves where the dashed lines are orthogonal vectors and the position of the control point p is placed along one of those vectors with different displacements Δp from the center point.

In the next sections, we present the unary potential which is used to define some probable coordinates for instrument parts, followed by different types of potential functions to impose kinematic constraints on the instrument parts and represent our prior model of the instrument.

3.1. Unary Potentials. The unary potential functions are designed to give a score for each instrument part hypothesis. Each hypothesis has a confidence value which is a probability assigned to the pixel in 2D images to express its degree of belonging to a specific instrument part. A regression forest [14] is trained on histogram of oriented gradients (HOG) [15] features for this purpose and regarded as a multiclass detector. The output of the regression forest is a class label prediction for each hypothesis and a confidence value. The number of class labels is set to the number of random variables in the CRF plus one for the background. The confidence value for each instrument part hypothesis is defined in

$$\Phi^{\text{Conf}}\left(y_i, x\right) = \frac{1}{T}\sum_{j=1}^{T}\pi_j\left(x\right), \tag{2}$$

where T is the number of trees in the forest and $\pi_j(x)$ is the probability assigned by one tree to y_i to express its belonging to a specific instrument part. The probability is given based on testing the features x associated with y_i. The term $\Phi^{\text{Temp}}(y_i, P_i)$ favors part hypotheses which are close to the last inferred part P_i based on the distance between them, as given by

$$\Phi^{\text{Temp}}\left(y_i, P_i\right) = \mathrm{e}^{-\|y_i - P_i\|_2^2/2}. \tag{3}$$

3.2. Binary Translation Potentials. The distance between the tips and the center point changes at different scales and orientations. The translation potentials model these translations of the left and the right tips to the center point by measuring the connectivity between the hypotheses of the instrument parts involved in the translational edges as shown in Figure 2(b). For example, given one hypothesis y_i of the left part and one hypothesis y_j of the center part detections, the connectivity between them is computed along different quadratic Bézier curves controlled by the position of the control point $P \in R^2$, as shown in Figure 3. The control point P is placed along the orthogonal vector to the vector (y_i, y_j). The distance of the point P to y_j specifies the shape of the curve connecting y_i and y_j. By denoting this curve as $C_{y_i}^{y_j}(P)$, the probabilistic connectivity along each curve is given by the following equation:

$$\text{Conn}\left(C_{y_i}^{y_j}\left(P\right)\right) = \frac{1}{k^2}\sum_{j=1}^{S}\left|s_j\right|^2, \tag{4}$$

in which k is a normalization factor. The curve is assumed to consist of $S \in R$ segments. Each segment s_j is a connected component of pixels along one curve. The connected components are extracted from the binary image created by thresholding the gradient image of the input microscopic image. The points y_i and y_j are overlaid on the binary image and considered strongly connected if at least one of Bézier curves aligned to the gripper edges curvature. This curve might consist of zero (not connected hypotheses where $C_{y_i}^{y_j}(P)$ is set to ε for numerical stability), one, or many segments. Changing the position of P by different Δp values enables the algorithm to handle various types of forceps with different curvatures along the gripper as shown in Figure 3. The connectivity measure in (4) is modeled to favor longer segments and penalize short ones in order to be robust in case of noisy images. The translation potential function keeps the maximum probability among all curves and it is defined in (5). A higher value of this probability means stronger connectivity and higher potential of the hypotheses to belong to the gripper end points:

$$\Psi^{\text{Conn}}\left(y_i, y_j\right) = \max_{\Delta p}\text{Conn}\left(C_{y_i}^{y_j}\left(P + \Delta p\right)\right). \tag{5}$$

The connectivity along the left and right parts of the gripper are calculated in the same way but with different positioning of the control point P.

3.3. Ternary Potentials. The relative length function Ψ^{RLen} is used to model the relative length between the left and right gripper parts as a Gaussian distribution and is given in (6). The function is designed to increase the algorithm robustness in case of false detections of structures like vessels near the instrument tips. The model parameters $\mu_{i,j,k}^{\text{RLen}}$ and $\sigma_{i,j,k}^{\text{RLen}}$ are estimated from the ground truth. Moreover, the gripper length should be consistent with shaft length in the ROI from which the configurations are selected. Hence, the

consistency function $\Psi^{\text{Cons}} \in \{1, \varepsilon\}$ is modeled to favor selected gripper parts with lengths less than half the size of the ROI side length. Otherwise, the output of the function is a small probability (ε) to penalize this configuration. In this way, the inconsistent combinations of parts hypotheses are penalized:

$$\Psi^{\text{RLen}}\left(y_i, y_j, y_k\right) = \mathcal{N}\left(\left(\left|y_i - y_j\right|, \left|y_i - y_k\right|\right) \mu_{i,j,k}^{\text{RLen}}, \sigma_{i,j,k}^{\text{RLen}}\right), \quad (6)$$

where y_i, y_j, and y_k are center, left, and right hypotheses, respectively.

3.4. Quaternary Rotation Potential. Any configuration y of the instrument forms an angles triple $\theta = \{\theta_i, i = 1, 2, 3\}$ among its parts treated as random variables. The rotation potential in (7) models the relations between these random variables as a mixture of two multivariate Gaussian distributions. One distribution models the relation among the variables when the instrument is closed or is about to be closed, while the other distribution is for the open instrument with different degrees. The parameters for each distribution (the mean $\mu_{i,j,k,l}^{R_n}$ and the covariance $\Sigma_{i,j,k,l}^{R_n}$) are estimated from the ground truth, where $n = 1$ for one distribution and $n = 2$ for the other:

$$\Psi^{\text{Rot}}\left(y_i, y_j, y_k, y_l\right) = \sum_{n=1}^{2} \mathcal{N}\left((\theta)_{i,j,k,l} \mid \mu_{i,j,k,l}^{R_n}, \Sigma_{i,j,k,l}^{R_n}\right), \quad (7)$$

where y_i, y_j, y_k, and y_l are left, center, right, and shaft hypotheses, respectively.

3.5. Inference of the Instrument Pose. We used genetic algorithms [16] to infer an approximate solution which maximizes the posterior equation as

$$\widehat{y} = \arg\max_{y} p\left(y \mid x, P\right). \quad (8)$$

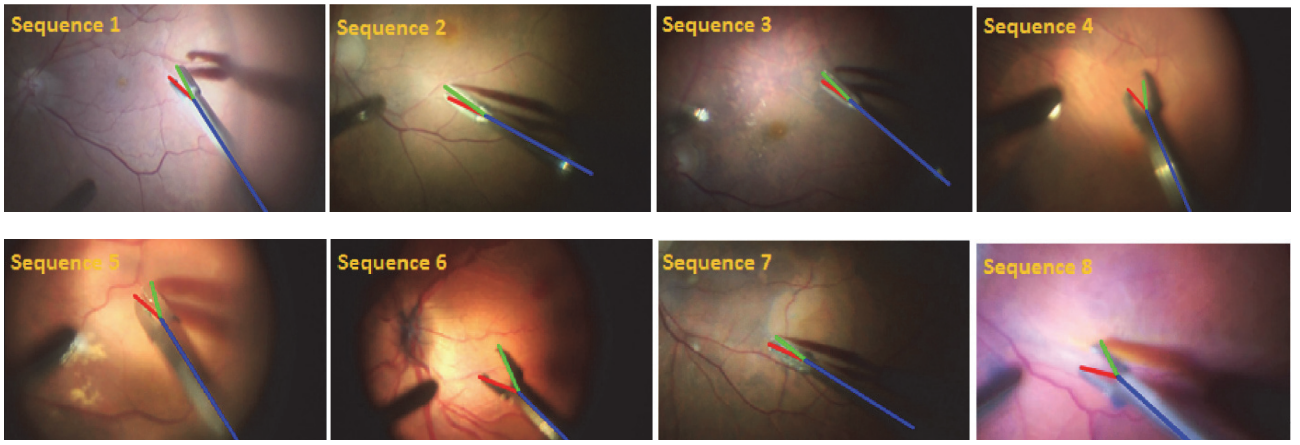
The most important parts of the genetic algorithms are the representation of the chromosomes and the definition of the fitness function. Each chromosome is represented by one configuration with four genes $\langle y_i, y_j, y_k, y_l \rangle$ representing the joints coordinates. The fitness function is set to the posterior function given in (1), which depends on the prior models $p(y)$ of the instrument kinematics and the initial hypotheses probabilities given by the regression forest. The algorithm starts by initial random generation of 1000 configurations which considered the initial population. Among those configurations, the crossover is applied pairwise by interleaving the genes at specific index to generate more variations from the current population. However, to enable the algorithm skipping local maxima during optimization, mutation operation is employed to replace random genes with others from the neighborhood. The produced configurations are evaluated using the fitness function, and a new generation is formed from the best evaluated configurations. The solution is obtained after a fixed number of iterations or no convergence in two successive generations.

Once the pose is estimated in the first frame, a reduced Region of Interest (ROI) is defined around the instrument center point to limit our detection space in the next frames. This ROI is expanded gradually when any instrument part is missing in the unary detections or when the confidence from the inferred pose is low. Low confidence of the final solution after optimization happens with either (1) low likelihood of the rotation distributions or (2) the consistency potential output being small (ε). These cases mean either that the solution cannot have the normal forceps shape or that it has been formed from false detections in ROI, which requires the reinitialization to be triggered automatically by expanding the ROI.

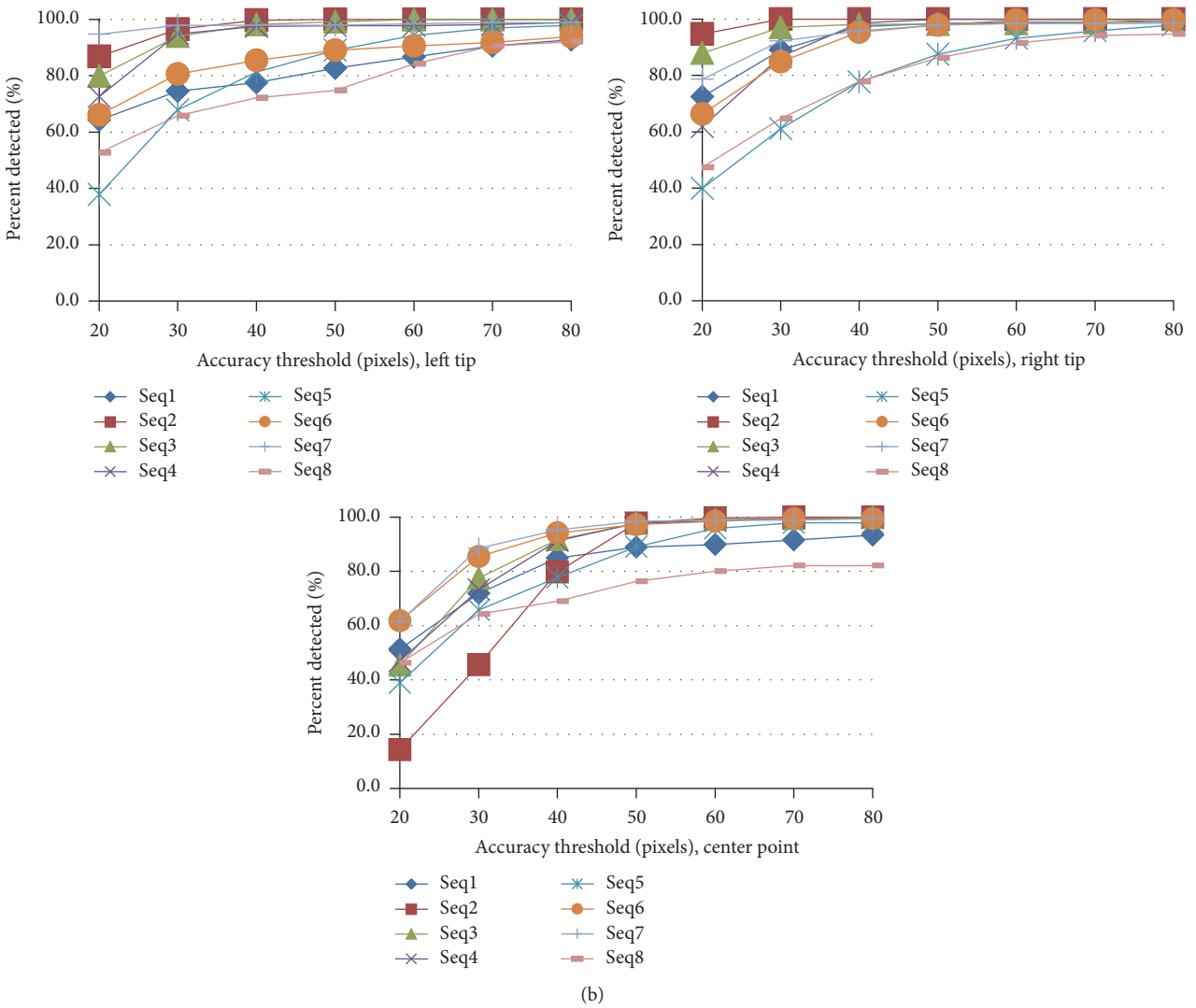
4. Experiments and Results

The experimental validation of the proposed method is carried out on three different microsurgery datasets. The first dataset, referred to as "Zeiss dataset," consists of eight sequences of surgeries performed on human eyes with frame resolution of 1920×1080 pixels, downsampled to one-fourth of the original size. Experiments on original size sequences prove the downsampling to have minimal effect on the detection accuracy. The second dataset is publicly available [10] with 1171 images of 640×480 pixels. No downsampling is performed on this dataset. The third dataset is a laparoscopic surgery dataset with 1000 images available on YouTube (http://www.youtube.com/watch?v=IVp1sgjQ5To). The proposed algorithm is evaluated by estimating the pose of one of the instruments present in the laparoscopic surgery since the other instrument has a fixed pose. The performance of the algorithm was evaluated using three different metrics: (1) accuracy threshold score defined by Sznitman et al. [10] to measure the pixel-wise detection accuracy for each instrument joint, (2) the strict Percentage of Correct Parts (strict PCP) [17] for gripper parts detection accuracy, and (3) the angular threshold score defined in [5] to measure the accuracy of estimating the shaft's orientation. The algorithm runs at 15-fps for public and laparoscopic datasets and 18-fps for Zeiss dataset on a normal personal computer. For the regression forest 50 trees with maximum depth of 25 are used. The HOG features bin size is set to 9 and the patch size is 50×50 pixels.

4.1. Zeiss Dataset. The algorithm was evaluated on 8 sequences as shown in Figure 4, where each sequence was taken from different surgery with different conditions. To achieve maximum reliability in clinical use, only 200 images from the first 4 sequences were used for training. The testing was done on the remaining images from each sequence in addition to 4 other unseen sequences. The number of testing images from each dataset is listed in Table 1. Each training frame has 4 annotated points: left and right tips, center point, and a point on the shaft centerline. 200 samples from the training images were manually clustered to open and close states to estimate the parameters of the rotation Gaussian distributions. Since the instrument shaft diameter is 50 pixels, we evaluate using values between 20 and 80 pixels for



(a)



(b)

Figure 4: (a) 8 samples from each sequence of Zeiss dataset with pose estimation; (b) the accuracy threshold for left, right, and center points, respectively.

Table 1: Strict PCP scores for $\alpha = 0.5$ on Zeiss dataset.

Zeiss sequences	Seq 1	Seq 2	Seq 3	Seq 4	Seq 5	Seq 6	Seq 7	Seq 8
#Testing images	590	400	400	400	200	400	200	200
Left PCP	91	99	98	98	92	85	96	75
Right PCP	93	99	99	99	93	94	97	76

Table 2: Strict PCP scores for $\alpha = 0.5$ on public and laparoscopic (Lap) datasets.

	Proposed				MC-15			
Public/Lap sequences	Seq 1	Seq 2	Full	Lap	Seq 1	Seq 2	Full	Lap
Left PCP	**97**	93	89	89	95	**97**	N/A	N/A
Right PCP	95	**95**	89	90	**97**	**95**	N/A	N/A

the accuracy threshold. Figure 4 shows the percentage of correctly predicted locations for different joints of the instrument. The results show that in 90% of the testing images the tips are detected with less than 50 pixels (the shaft diameter) error. The strict PCP scores of the left and right gripper's parts for $\alpha = 0.5$ (which used for human pose estimation evaluation) for each sequence are depicted in Table 1 which show the robustness of the algorithm and its ability to generalize to new sequences.

4.2. Public Dataset. The proposed method was compared with the state-of-the art methods: MC-15 [13], MC-14 [2], MC-12 [10], SCV [18], MI [19], and SSD. The evaluation includes two sequences of the public datasets. The third sequence is omitted, as in [2], due to its short length which makes it ill-suited for training purposes. In the first experiment, the training is done separately on the first half of each sequence and testing was on the second half. The detection accuracy of the center point is shown in Figure 5 which shows comparability of the proposed method to the state-of-the-art methods with the advantage of not requiring manual reinitialization. For example, at threshold of 20 pixels (the shaft diameter), the center points are detected correctly in more than 95% of the images in both cases. The accuracy threshold scores for detecting the two tips of the forceps in each sequence are depicted in Figure 7.

In the second experiment, the training is performed on the full dataset (the first two halves of the two sequences together) and the testing is done on the second halves. The performance of detecting forceps tips and forceps center point is shown in Figure 7 labeled with the prefix full. The strict PCP scores for both experiments are listed in Table 2 and compared to MC-15 [13] which is the only state-of-the-art method that can locate the forceps tips even though it is only tracking method and uses manual initialization to handle tracking failures in live surgery.

4.3. Laparoscopic Dataset. We compared our performance with MC-15 [13], MC-12 [10], ITOL [11], MF [11], and DT [11]. Similar to these methods, training was done on the first half of the dataset and the testing on the second half. Comparing the performance of our method in detecting the center point with the other methods using accuracy threshold is shown in Figure 6. It is obvious that our method outperforms most state-of-the-art methods and achieves similar results to ITOL which is also a tracking method and impractical for live surgery due to the required manual reinitialization. The accuracy threshold scores of detecting each tip are shown in Figure 7 while all other methods do not detect them in this challenging dataset. The PCP scores are given in Table 2 which show even high detection accuracy of both gripper's parts.

Figures 8(a) and 8(b) show the performance of our algorithm to estimate the orientation of the shaft while varying the angular threshold from 3 to 24 deg. It is evident that in 85 percent of the images, the orientation is detected with deviation less than 15 deg.

5. Results Discussion

The proposed approach shows high accuracy of instrument joints localization in real-time performance. This accuracy is attributed to modeling the dependencies between instrument parts as CRF model, while other methods do not consider these dependencies and rely only on individual parts detection. These dependencies are built on top of random forest outputs trained using only gradient-based (HOG) features to serve as unary detections functions. Unlike other intensity-based tracker methods, relying on HOG features makes our approach robust enough to illumination changes during surgery. Moreover, it reduces the amount of training samples needed for training large changes in instrument appearance. This is why, in the first dataset, our algorithm needs only 200 samples from only 4 sequence and it is able to run on testing images with 3 times the size of the training ones. Practically, it can run on even longer sequences since there is no need to train more samples to account for new illumination changes. This has been proven by running the algorithm on 4 other unseen sequences and achieving high performance which is considered a great achievement of our approach in comparison with the state-of-the-art methods MC-15 [13] and ITOL [11]. Moreover, relying on detected structural parts

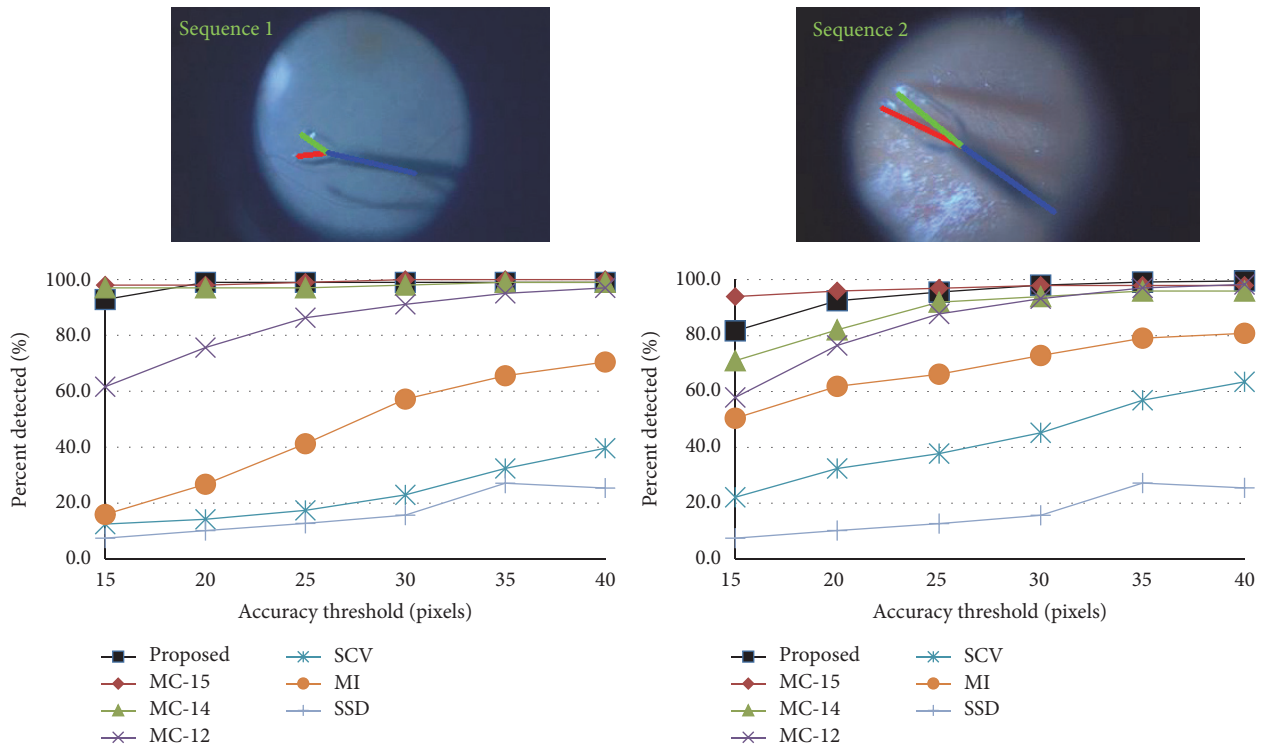


Figure 5: Threshold accuracy for each of the public sequences separately.

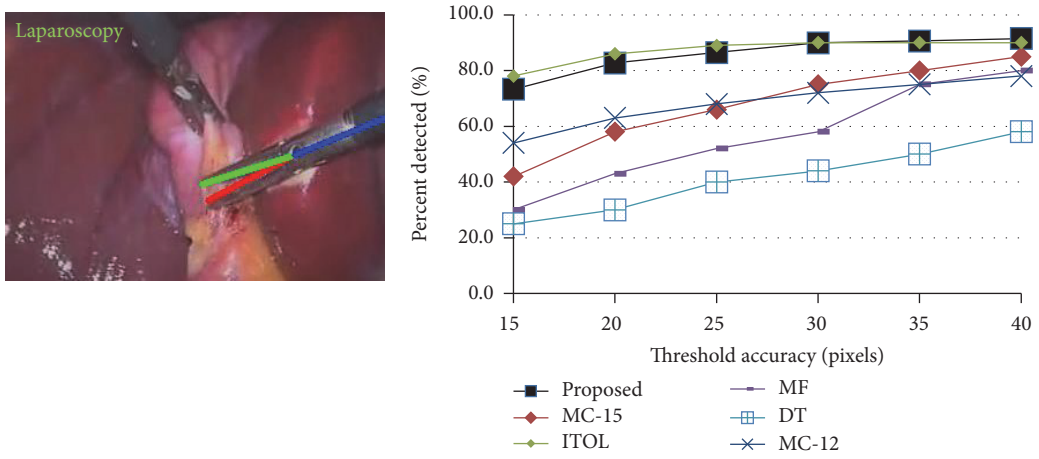


Figure 6: Threshold accuracy for laparoscopic dataset.

using HOG features bring new advantage to our method which is being able to sense some confidence signals. This feedback is employed for automatic recovery process, which is missing in most other methods, to localize again the instrument after its disappearance without surgeon's intervention. The results also presented high PCP scores on most of the retinal sequences. However, in sequence 8, the PCP score is not as high as the other sequences due to the blurriness of the images which makes the detection of the gripper edges very difficult. Hence, the connectivity potential function will not be able to give fair preferences to some configurations. Coming to the public dataset, PCP scores of our method show comparable results to MC-15 [13]. However, the advantage of our approach is the ability to work without stopping on these sequences, while in sequence 2, MC-15 [13] needed the manual reinitialization twice to handle instrument disappearance from the scene. Comparing on laparoscopic dataset, our approach outperforms MC-15 [13] by at least 20% at most of the accuracy thresholds in localizing the instrument center point as shown in Figure 6 and achieves very close performance to ITOL [11]. However, ITOL cannot detect the forceps two tips as well as it is just intensity-based

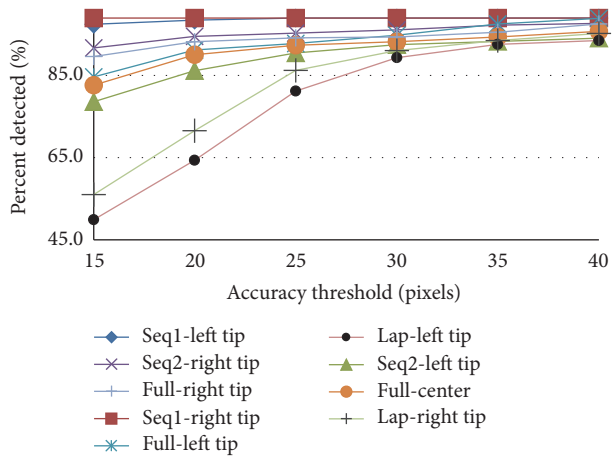


Figure 7: Accuracy threshold for different forceps joints of the public (full and separate sequences) and laparoscopic (Lap) datasets.

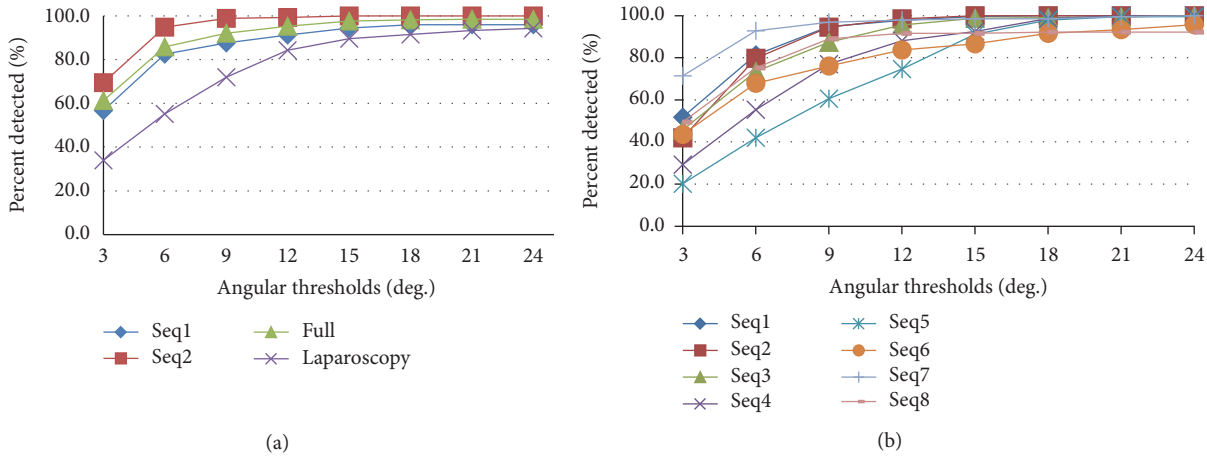


Figure 8: (a) Angular threshold scores for public and laparoscopic sequences; (b) angular threshold scores for Zeiss sequences.

tracking algorithm. Hence, our algorithm tends to be more robust and practical for real surgeries due to its ability to localize the instrument left and right tips with high accuracy.

One more important strength point of the proposed approach is the ability to estimate the orientation of the instrument shaft. Unlike other approaches, the orientation is treated as a part in our CRF model, and this characteristic makes our approach successful one for the full integration with OCT imaging to position OCT scans according to given coordinates and orientation. The angular threshold results show also high accuracy in estimating the instrument orientation in all sequences of the different datasets.

6. Conclusions

We presented a new approach for localizing the forceps tips and center point as well as estimating the orientation of its shaft. The approach models the instrument detection, tracking, and pose estimation as a CRF's inference problem. The performance of the proposed approach has been evaluated on retinal and laparoscopic surgeries using three different metrics. The algorithm generates all parameters needed for OCT device in order to position OCT scans automatically in real surgery. It also achieves real-time performance and works on real surgery sequences. Moreover, it does not require manual initialization since it tracks the instrument by constantly detecting its parts and maintains a confidence value to reinitialize the detection automatically whenever it is needed. The method demonstrates high detection rate of the instrument joints on long sequences as well as comparable results to the state-of-the-art methods without the need of manual reinitialization.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Mohamed Alsheakhali and Abouzar Eslami contributed equally to this work.

Acknowledgments

This work was supported by the German Research Foundation (DFG) and the Technische Universität München within the Open Access Publishing Funding Programme.

References

[1] H. Roodaki, K. Filippatos, A. Eslami, and N. Navab, "Introducing augmented reality to optical coherence tomography in ophthalmic microsurgery," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR '15)*, pp. 1–6, Fukuoka, Japan, September 2015.

[2] R. Sznitman, C. Becker, and P. Fua, "Fast part-based classification for instrument detection in minimally invasive surgery," in *Proceedings of the 17th International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI '14)*, pp. 692–699, Boston, Mass, USA, September 2014.

[3] M. Alsheakhali, A. Eslami, and N. Navab, "Detection of articulated instruments in retinal microsurgery," in *Proceedings of the International Symposium on Biomedical Imaging (ISBI '16)*, pp. 107–110, Prague, Czech Republic, April 2016.

[4] Z. Pezzementi, S. Voros, and G. D. Hager, "Articulated object tracking by rendering consistent appearance parts," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '09)*, pp. 3940–3947, IEEE, Kobe, Japan, May 2009.

[5] M. Alsheakhali, M. Yigitsoy, A. Eslami, and N. Navab, "Surgical tool detection and tracking in retinal microsurgery," in *Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9415 of *Proceedings of SPIE*, Orlando, Fla, USA, February 2015.

[6] A. Reiter, P. K. Allen, and T. Zhao, "Feature classification for tracking articulated surgical tools," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds., vol. 7511 of *Lecture Notes in Computer Science*, pp. 592–600, Springer, Berlin, Germany, 2012.

[7] R. Wolf, J. Duchateau, P. Cinquin, and S. Voros, "3d tracking of laparoscopic instruments using statistical and geometric modeling," in *Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI '11)*, pp. 203–210, Toronto, Canada, September 2011.

[8] C.-J. Chen, W. S.-W. Huang, and K.-T. Song, "Image tracking of laparoscopic instrument using spiking neural networks," in *Proceedings of the 13th International Conference on Control, Automation and Systems (ICCAS '13)*, pp. 951–955, Seoul, Republic of Korea, October 2013.

[9] R. Sznitman, A. Basu, R. Richa et al., "Unified detection and tracking in retinal microsurgery," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2011: 14th International Conference, Toronto, Canada, September 18–22, 2011, Proceedings, Part I*, vol. 6891 of *Lecture Notes in Computer Science*, pp. 1–8, Springer, Berlin, Germany, 2011.

[10] R. Sznitman, K. Ali, R. Richa, R. H. Taylor, G. D. Hager, and P. Fua, "Data-driven visual tracking in retinal microsurgery," in *Medical Image Computing and Computer-Assisted Intervention —MICCAI 2012: 15th International Conference, Nice, France, October 1–5, 2012, Proceedings, Part II*, vol. 7511 of *Lecture Notes in Computer Science*, pp. 568–575, Springer, Berlin, Germany, 2012.

[11] Y. Li, C. Chen, X. Huang, and J. Huang, "Instrument tracking via online learning in retinal microsurgery," in *Proceedings of the 17th International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI '14)*, pp. 464–471, Boston, Mass, USA, September 2014.

[12] M. Allan, P.-L. Chang, S. Ourselin et al., "Image based surgical instrument pose estimation with multi-class labelling and optical flow," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part I*, vol. 9349 of *Lecture Notes in Computer Science*, pp. 331–338, Springer, Berlin, Germany, 2015.

[13] N. Rieke, D. J. Tan, M. Alsheakhali et al., "Surgical tool tracking and pose estimation in retinal microsurgery," in *Proceedings of the 18th International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI '15)*, vol. 9349, pp. 266–273, Springer, Munich, Germany, 2015.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 886–893, IEEE, San Diego, Calif, USA, June 2005.

[16] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2nd edition, 2002.

[17] V. Ferrari, M. Marín-Jiménez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[18] M. R. Pickering, A. A. Muhit, J. M. Scarvell, and P. N. Smith, "A new multi-modal similarity measure for fast gradient-based 2D-3D image registration," in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '09)*, pp. 5821–5824, September 2009.

[19] M. Balicki, E. Meisner, R. Sznitman, R. Taylor, and G. Hager, *Visual Tracking of Surgical Tools for Proximity Detection in Retinal Surgery*, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 6584725, 16 pages
http://dx.doi.org/10.1155/2016/6584725

Research Article

Exploring Deep Learning and Transfer Learning for Colonic Polyp Classification

Eduardo Ribeiro,[1,2] Andreas Uhl,[1] Georg Wimmer,[1] and Michael Häfner[3]

[1]*Department of Computer Sciences, University of Salzburg, Salzburg, Austria*
[2]*Department of Computer Sciences, Federal University of Tocantins, Palmas, TO, Brazil*
[3]*St. Elisabeth Hospital, Vienna, Austria*

Correspondence should be addressed to Eduardo Ribeiro; ufg.eduardo@gmail.com

Received 10 August 2016; Accepted 4 October 2016

Academic Editor: Ayman El-Baz

Recently, Deep Learning, especially through Convolutional Neural Networks (CNNs) has been widely used to enable the extraction of highly representative features. This is done among the network layers by filtering, selecting, and using these features in the last fully connected layers for pattern classification. However, CNN training for automated endoscopic image classification still provides a challenge due to the lack of large and publicly available annotated databases. In this work we explore Deep Learning for the automated classification of colonic polyps using different configurations for training CNNs from scratch (or full training) and distinct architectures of pretrained CNNs tested on 8-HD-endoscopic image databases acquired using different modalities. We compare our results with some commonly used features for colonic polyp classification and the good results suggest that features learned by CNNs trained from scratch and the "off-the-shelf" CNNs features can be highly relevant for automated classification of colonic polyps. Moreover, we also show that the combination of classical features and "off-the-shelf" CNNs features can be a good approach to further improve the results.

1. Introduction

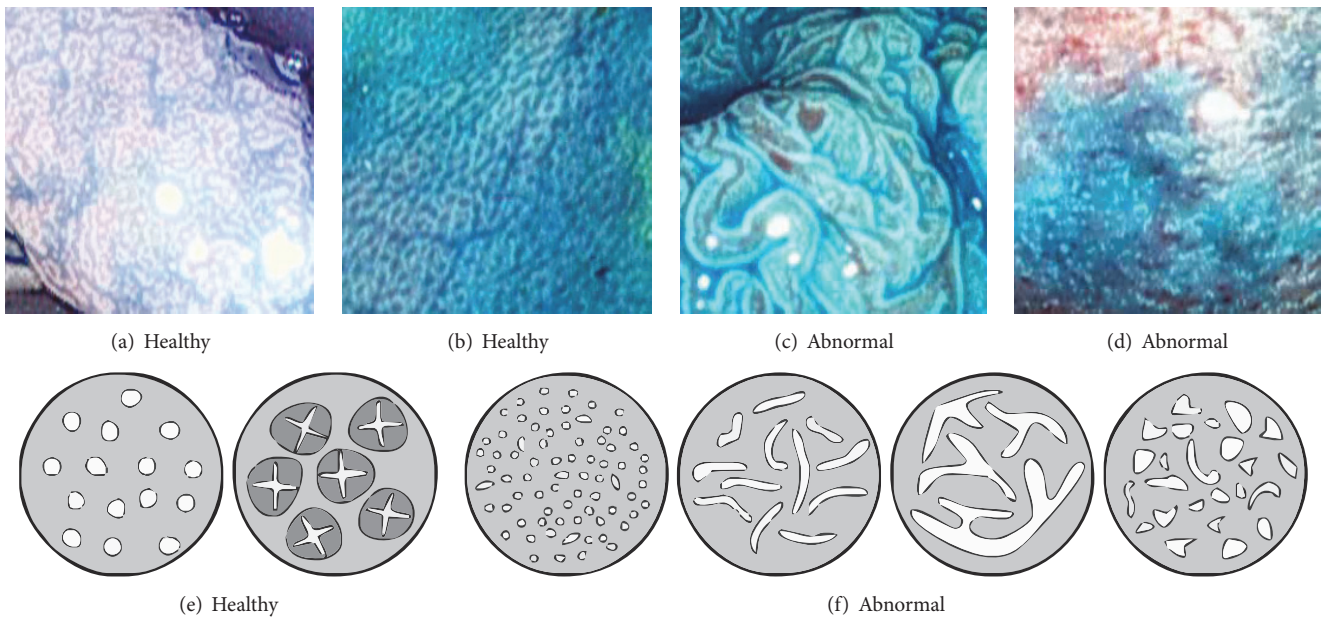
The leading cause of deaths related to the intestinal tract is the development of cancer cells (polyps) in its many parts. An early detection (when the cancer is still at an early stage) and a regular exam to everyone over an age of 50 years can reduce the risk of mortality among these patients. More specifically, colonic polyps (benign tumors or growths which arise on the inner colon surface) have a high occurrence and are known to be precursors of colon cancer development.

Endoscopy is the most common method for identifying colon polyps and several studies have shown that automatic detection of image regions which may contain polyps within the colon can be used to assist specialists in order to decrease the polyp miss rate [1, 2].

The automatic detection of polyps in a computer-aided diagnosis (CAD) system is usually performed through a statistical analysis based on color, shape, texture, or spatial features applied to the videos frames [3–6]. The main problems for the detection are the different aspects of color, shape, and textures of polyps, being influenced, for example, by the viewing angle, the distance from the capturing camera, or even by the colon insufflation as well as the degree of colon muscular contraction [5].

After detection, the colonic polyps can be classified into three different categories: hyperplasic, adenomatous, and malignant. Kudo et al. [7] proposed the so-called "pit-pattern" scheme to help in diagnosing tumorous lesions once suspicious areas have been detected. In this scheme, the mucosal surface of the colon can be classified into 5 different types designating the size, shape, and distribution of the pit structure [8, 9].

As can be seen in the Figures 1(a)–1(d), these five patterns also allow the division of the lesions into two main classes: (1) normal mucosa or hyperplastic polyps (healthy class) and (2) neoplastic, adenomatous, or carcinomatous structures



(a) Healthy (b) Healthy (c) Abnormal (d) Abnormal

(e) Healthy (f) Abnormal

Figure 1: Example images of the two classes (a–d) and the pit-pattern types of these two classes (e–f).

(abnormal class). This approach is quite relevant in clinical practice as shown in a study by Kato et al. [10].

In the literature, existing computer-aided diagnosis techniques generally make use of feature extraction methods of color, shape, and texture in combination with machine learning classifiers to perform the classification of colon polyps [9, 11, 12]. For example, the dual-tree complex wavelet transform DT-CWT features proved to be quite suitable for the distinction of different types of polyps as can be seen in many works like, for example, [13–15]. Other features were also proved to be quite suitable for colonic polyp classification as the Gabor wavelets [16], vascularization features [17], and directional wavelet transform features [18]. Particularly, in the work of Wimmer et al. [18], using the same 8 colonic polyp databases of this work, an average accuracy of 80.3% was achieved in the best scenario. In this work, we achieve an average accuracy of 93.55% in our best scenario.

The main difficulty of the feature extraction methods is the proper characterization of these patterns due to several factors as the lack or excess of illumination, the blurring due to movement or water injection, and the appearance of polyps [5, 9]. Also, to find a robust and a global feature extractor that summarizes and represents all these pit-pattern structures in a single vector is very difficult and Deep Learning can be a good alternative to surpass these problems. In this work we explore the use of Deep Learning through Convolutional Neural Networks (CNNs) to develop a model for robust feature extraction and efficient colonic polyp classification.

To achieve this, we test the use of CNNs trained from scratch (or full training) and off-the-shelf CNNs (or pre-trained) using them as medical imaging feature extractors. In the case of the CNN full training we assume that a feature extractor is formed during the CNN training, adapting to the context of the database and particularly in the case of off-the-shelf CNNs we consider that the patterns learned in the original database can be used in colonoscopy images for colonic polyp classification. In particular, we explore two different architectures for the training from scratch and six different off-the-shelf architectures, describing and analyzing the effects of CNNs in different acquisition modes of colonoscopy images (8 different databases). This study was motivated by recent studies in computer vision addressing the emerging technique of Deep Learning presented in the next section.

2. Materials and Methods

2.1. Using CNNs on Small Datasets. Some researchers propose replacing handcrafted feature extraction algorithms with Deep Learning approaches that act as features extractor and image classifier at the same time [19]. For example, the Deep Learning approach using CNNs takes advantage of many consecutive convolutional layers followed by pooling layers to reduce the data dimensionality making it, concomitantly, invariant to geometric transformations. Such convolution filters (kernels) are built to act as feature extractors during the training process and recent research indicates that a satisfactorily trained CNN with a large database can perform properly when it is applied to other databases, which can mean that the kernels can turn into a universal feature extractor [19]. Also, Convolutional Neural Networks (CNNs) have been demonstrated to be effective for discriminative pattern recognition in big data and in real-world problems, mainly to learn both the global and local structures of images [20].

Many strategies exploiting CNNs can be used for medical image classification. These strategies can be employed according to the intrinsic characteristics of each database [21] and two of them, mostly used when it comes to CNN training, are described in the following part.

When the available training database is large enough, diverse, and very different from the database used in all the available pretrained CNNs (in a case of transfer learning), the most appropriate approach would be to initialize the CNN weights randomly (training the *CNN trained from scratch*) and train it according to the medical image database for the kernels domain adaptation, that is, to find the best way to extract the features of the data in order to classify the images properly. The main advantage of this approach is that the same method can be used for the extraction of strong features that are invariant to distortion and position at the same time of the image classification. Finally, the Neural Network Classifier can make use of these inputs to delineate more accurate hyperplanes helping the generalization of the network.

This strategy, although ideal, is not widely used due to the lack of large and annotated medical image database publicly available for training the CNN. However, some techniques can assist the CNN training from scratch with small datasets and the most used approach is data augmentation. Basically, in data augmentation, transformations are applied to the image making new versions of it to increase the number of samples in the database. These transformations can be applied in both the training and the testing phase and can use different strategies such as cropping (overlapped or not), rotation, translation, and flipping [22]. Experiments show that using these techniques can be effective to combat overfitting in the CNN training and improve the recognition and classification accuracy [22, 23].

Furthermore, when the database is small, the best alternative is to use an *off-the-shelf CNN* [21]. In this case, using a pretrained CNN, the last or next-to-last linear fully connected layer is removed and the remaining pretrained CNN is used as a feature extractor to generate a feature vector for each input image from a different database. These feature vectors can be used to train a new classifier (such as a support vector machine, SVM) to classify the images correctly. If the original database is similar to the target database, the probability that the high-level features describe the image correctly is high and relevant to this new database. If the target database is not so similar to the original, it can be more appropriate to use higher-level features, that is, features from previous layers of CNN.

In this work, besides using a CNNs trained from scratch, we consider the knowledge transfer between natural images and medical images using off-the-shelf pretrained CNNs. The CNN will project the target database samples into a vector space where the classes are more likely to be separable. This strategy was inspired by the work of Oquab et al. [24], which uses a pretrained CNN on a large database (ImageNet) to classify images in a smaller database (Pascal VOC dataset) with improved results. Unlike that work, rather than copy the weights of the original pretrained CNN to the target CNN with additional layers, we use the pretrained CNN to project data into a new feature space through the propagation of the colonic polyp database into the CNN getting the resultant vector from the last CNNs layer, obtaining a new representation for each input sample. Subsequently, we use the feature vector set to train a linear classifier (e.g., support vector machines) in this representation to evaluate the results as used in [25, 26].

2.2. CNNs and Medical Imaging. In recent years there has been an increased interest in machine learning techniques that is based not on hand-engineered feature extractors but using raw data to learn the representations [19].

Among the development of efficient parallel solvers together with GPUS, the use of Deep Learning has been extensively explored in the last years in different fields of application. Deep Learning is intimately related to the use of raw data to do high-level representations of this knowledge through a large volume of annotated data. However, when it comes to the medical area, this type of application is limited by the problem of the lack of large, annotated, and publicly available medical image databases such as the existing natural image databases. Additionally, it is a difficult and costly task to acquire and annotate such images and due to the specific nature of different medical imaging modalities which seems to have different properties according to each modality the situation is even aggravated [21, 27].

Recently, works addressing the use of Deep Learning techniques in medical imaging have been explored in many different ways mainly using CNNs trained from scratch. In biomedical applications, examples include mitosis detection in digital breast cancer histology [28] and neuronal segmentation of membranes in electron microscopy [29]. In Computer-Aided Detection systems (CADe systems), examples include a CADe of pulmonary embolism [30], computer-aided anatomy detection in CT volumes [31], lesion detection in endoscopic images [32], detection of sclerotic spine metastases [33], and automatic detection of polyps in colonoscopy videos [27, 34, 35]. In medical image classification, CNNs are used for histopathological image classification [36], digestive organs classification in wireless capsule endoscopy images [37, 38], and automatic colonic polyp classification [39]. Besides that, CNNs have also been explored to improve the accuracy of CADe systems knee cartilage segmentation using triplanar CNNs [40].

Other recent studies show the potential for knowledge transfer from natural images to the medical imaging domain using off-the-shelf CNNs. Examples include the identification and pathology of X-ray and computer tomography modalities [25], automatic classification of pulmonary perifissural nodules [41], pulmonary nodule detection [26], and mammography mass lesion classification [42]. Moreover, in [26], Van Ginneken et al. show that the combination of CNNs features and classical features for pulmonary nodule detection can improve the performance of the model.

2.2.1. CNNs Trained from Scratch: Architecture. In this section we briefly describe the components of a CNN and how it can be used to perform the CNN from scratch.

A CNN is very similar to traditional Neural Networks in the sense of being constructed by neurons with their respective weights, biases, and activation functions. The structure is basically formed by a sequence of convolution and pooling layers ending in a fully connected Neural Network as shown in Figure 2. Generally, the input of a CNN is $m \times m \times d$

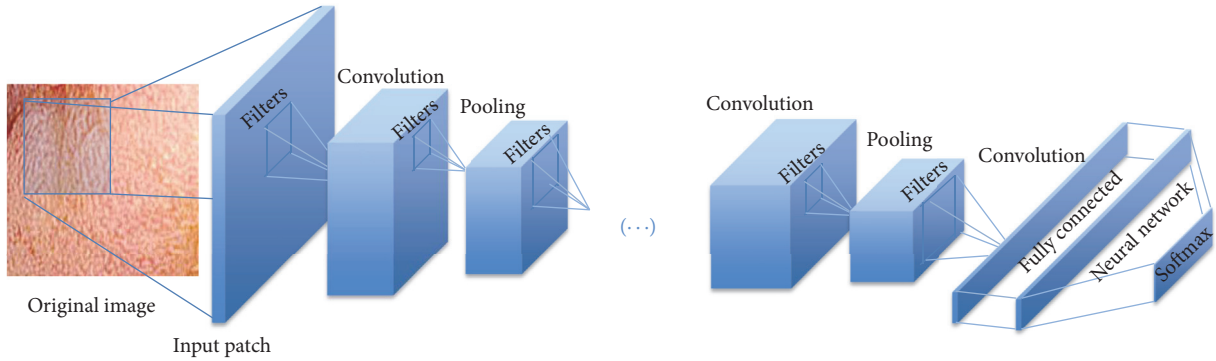


Figure 2: An illustration of the CNN architecture for colonic polyp classification.

image (or patch), where $m \times m$ is the dimension of the image and d is the number of channels (depth) of the image. The convolutional layer consists of k learnable filters (also called kernels) with size $n \times n \times d$ where $n \leq m$ which are convolved with the input image resulting in the so-called activation maps or feature maps. As classic Neural Networks, the convolution layer outputs are submitted to an activation function, for example, the ReLU rectifier function $f(x) = \max(0, x)$, where x is the neuron input. After the convolution, a pooling layer is included to subsample the image by average functions (mean) or max-pooling over regions of size $p \times p$. These functions are used to reduce the dimensionality of the data in the following layers (upper layers) and to provide a form of invariance to translation thus making overfitting control. In the convolution and pooling layers the stride has to be specified; the larger the stride, the smaller the overlapping, decreasing the output volume dimensions.

At the end of the CNN there is a fully connected layer as a regular Multilayer Neural Network with the Softmax function that generates a well-formed probability distribution on the outputs. After a supervised training, the CNN is ready to be used as a classifier or as a feature extractor in the case of transfer learning.

2.2.2. CNNs and Transfer Learning. Transfer learning is a technique used to improve the performance of machine learning by harnessing the knowledge obtained by another task. According to Pan and Yang [43], transfer learning can be defined by the following model. We give a domain D having two components: a feature space $X = \{x_1, x_2, \ldots, x_n\}$ and a probabilistic distribution $P(X)$; that is, $D = \{X, P(X)\}$. Also, we give a task T with two components: a ground truth $Y = \{y_1, y_2, \ldots, y_n\}$ and an objective function $T = \{Y, f(\cdot)\}$ assuming that this function can be learned through a training database. Function $f(\cdot)$ can be used to predict the correspondent class $f(x)$ of a new instance x. From a probabilistic point of view, $f(x)$ can be written as $P(y \mid x)$. In colonic polyp classification, usually, a feature extractor is used to generate the feature space. A given training database X associated to the ground truth Y consisting of the pairs $\{x_i, y_i\}$ is used to train and "learn" the function $f(\cdot)$ or $P(y \mid x)$ until it reaches a defined and acceptable error rate between the result of the function $f(x)$ and the ground truth Y.

In case of transfer learning, given a source domain $D_S = \{(x_{S_1}, y_{S_1}), (x_{S_2}, y_{S_2}), \ldots, (x_{S_n}, y_{S_n})\}$ and the learning task T_S and the target domain $D_T = \{(x_{T_1}, y_{T_1}), (x_{T_2}, y_{T_2}), \ldots, (x_{T_m}, y_{T_m})\}$ and the learning task T_T, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ using the knowledge in D_S and T_S, where $D_T \neq D_S$ and $T_T \neq T_S$.

Among the various categories of transfer learning, one, called inductive transfer learning, has been used with success in the pattern recognition area. In the inductive transfer learning approach an annotated database is necessary for the source domain as well as for the target domain. In this work, we apply transfer learning between two very different tasks using different labels ($Y_T \neq Y_S$) and different distributions ($P(Y_T \mid X_T) \neq P(Y_S \mid X_S)$). To bypass the difference between the probability distribution of the images $P(X_S)$, the last layer from the original function $f_S(\cdot)$ directly connected to the classification is removed being replaced by other linear function (as SVM) to adapt it to the new task T_T turning into the function $f_T(\cdot)$. In the following sections the functions $f_S(\cdot)$ used in this work are presented. Also, the use of transfer learning using pretrained CNNs can help to avoid the problem of lack of data in the medical field. The works of Razavian et al. [19] and Oquab et al. [24] suggest that the use of CNNs intermediate layer outputs can be used as input features to train other classifiers (such as support vector machines) for a number of other applications different from the original CNN obtaining a good performance.

Despite the difference between natural and medical images, some feature descriptors designed especially for natural images are used successfully in medical image detection and classification, for example, texture-based polyp detection [3], Fourier and Wavelet filters for colon classification [18], shape descriptors [44], and local fractal dimension [45] for colonic polyp classification. Additionally, recent studies show the potential of the knowledge transfer between natural and medical images using pretrained (off-the-shelf) CNNs [34, 46].

2.3. Experimental Setup

2.3.1. Data. The use of an integrated endoscopic apparatus with high-resolution acquisition devices has been an

Table 1: Number of images and patients per class of the CC-i-Scan databases gathered with and without CC (staining) and computed virtual chromoendoscopy (CVC).

i-Scan mode	No staining				Staining			
	¬CVC	i-Scan1	i-Scan2	i-Scan3	¬CVC	i-Scan1	i-Scan2	i-Scan3
Non-neoplastic								
Number of images	39	25	20	31	42	53	32	31
Number of patients	21	18	15	15	26	31	23	19
Neoplastic								
Number of images	73	75	69	71	68	73	62	54
Number of patients	55	56	55	55	52	55	52	47
Total number of images	112	100	89	102	110	126	94	85

important object of research in clinical decision support system area. With high-magnification colonoscopies it is possible to acquire images up to 150-fold magnified, revealing the fine surface structure of the mucosa as well as small lesions. Recent work related to classification of colonic polyps used highly-detailed endoscopic images in combination with different technologies divided into three categories: high-definition endoscope (with or without staining the mucosa) combined with the i-Scan technology (1, 2, and 3) [18], high-magnification chromoendoscopy [8], and high-magnification endoscopy combined with narrow band imaging [47].

Specifically, the i-Scan technology (Pentax) used in this work is an image processing technology consisting of the combination of surface enhancement and contrast enhancement aiming to help detect dysplastic areas and to accentuate mucosal surfaces and applying postprocessing to the reflected light being called virtual chromoendoscopy (CVC) [44].

There are three i-Scan modes available: i-Scan1, which includes surface enhancement and contrast enhancement, i-Scan2 that includes surface enhancement, contrast enhancement, and tone enhancement, and i-Scan3 that, besides including surface, contrast, and tone enhancement, increases lighting emphasizing the features of vascular visualization [18]. In this work we use an endoscopic image database (CC-i-Scan Database) with 8 different imaging modalities acquired by an HD endoscope (Pentax HiLINE HD+ 90i Colonoscope) with images of size 256 × 256 extracted from video frames either using the i-Scan technology or without any computer virtual chromoendoscopy (¬CVC).

Table 1 shows the number of images and patients per class in the different i-Scan modes. The mucosa is either stained or not stained. Despite the fact that the frames were originally in high-definition, the image size was chosen (i) to be large enough to describe a polyp and (ii) small enough to cover just one class of mucosa type (only healthy or only abnormal area). The image labels (ground truth) were provided according to their histological diagnosis.

2.3.2. Employed CNN Techniques. Due to the limitation of colonic polyp images to train a good CAD system from scratch, the main elements of the proposed method are defined in order to (1) extract and preprocess images aiming to have a database with a suitable size, (2) use CNNs for learning representative features with good generalization, and (3) enable the use of methods to avoid overfitting in the training phase.

To test the application of a CNN trained from scratch we used the i-Scan1 database without chromoscopy (staining the mucosa) that presents a good performance in the tests using classical features and pretrained CNNs (on average) and subsequently applying the best configuration to the i-Scan3 without chromoscopy database that presented the best results among the classical features results.

In the first experiment of CNN full training, it is proposed that an architecture should be trained with subimages of size 227 × 227 × 3 based on the work of [20] to fit into the chosen architecture. Usually, some simple preprocessing techniques are necessary for the image feature generation. In this experiment we apply normalization by subtracting the mean and dividing by the standard deviation of its elements as in [48] corresponding to local brightness and normalization contrast. We also perform data augmentation by flipping each original image horizontally and vertically and rotating the original image 90° to the right and left. Besides that, we flipped horizontally the rotated images, and then we flipped vertically the horizontally flipped image, totalizing 7 new samples for each original image. After the data augmentation (resulting in 800 images), we randomly extract 75 subimages of size 227×227×3 from each healthy image and 25 subimages from each abnormal image for the training set to balance the number of images in each class.

Also, in this experiment, to be able to compare the different architectures in a faster way, we used cross-validation evaluation with 10 different CNNs for each architecture. In nine of them, we removed 56 patients for training and used 6 for tests and, in one of them, we removed 54 patients for training and used 8 for test to assure that all the 62 patients are tested. The accuracy result given for each architecture is the average accuracy from each of the 10 CNNs trained based on the final classification of each image between the two classes.

For the second experiment in the CNN full training we propose to extract subimages of size 128 × 128 from the original images using the same approach as in the first experiment. In this case, we explore the hypothesis that the colonic polyp classification with the CNN can be done only with a part of the image, and then we trained the network with smaller subimages instead of the entire image. This helps to

reduce the size of the network reducing its complexity and can allow different polyp classifications in the same image using different subimages in different parts of the image. Additionally, choosing smaller regions in a textured image can diminish the degree of intraimage variances in the dataset as the neighborhood is limited.

Besides the different architectures for the training from scratch, we mainly explore six different off-the-shelf CNN architectures trained to perform classification on the ImageNet ILSVRC challenge data. The input of all tested pretrained CNNs has size of $224 \times 224 \times 3$ and the descriptions as well as the details of each CNN are given as follows:

(i) The *CNN VGG-VD* [49] uses a large number of layers with very small filters (3×3) divided into two architectures according to the number of their layers. The CNN *VGG-VD16* has 16 convolution layers and five pooling layers while the CNN *VGG-VD19* has 19 convolution layers, adding one more convolutional layer in three last sequences of convolutional layers. The fully connected layers have 4096 neurons followed by a Softmax classifier with 1000 neurons corresponding to the number of classes in the ILSVRC classification. All the layers are followed by a rectifier linear unit (ReLU) layer to induce the sparsity in the hidden units and reduce the gradient vanishing problem.

(ii) The *CNN-F* (also called Fast CNN) [22] is similar to the CNN used by Alex et al. [20] with 5 convolutional layers. The input image size is 224×224 and the fast processing is granted by the stride of 4 pixels in the first convolutional layer. The fully connected layers also have 4096 neurons as the CNN VGG-VD. Besides the original implementation, in this work, we also used the MatConvNet implementation (beta17 [50]) of this architecture trained with batch normalization and minor differences in its default hyperparameters and called here *CNN-F MCN*.

(iii) The *CNN-M* architecture (Medium CNN) [22] also has 5 convolutional layers and 3 pooling layers. The number of filters is higher than the Fast CNN: 96 instead of 64 filters in the first convolution layer with a smaller size. We also use the MatConvNet implementation called *CNN-M MCN*.

(iv) The *CNN-S* (Slow CNN) [22] is related to the "accurate" network from the Overfeat package [51] and also has smaller filters with a stride of 2 pixels in the first convolutional layer. We also use the MatConvNet implementation called *CNN-S MCN*.

(v) The *AlexNet* CNN [20] has five convolutional layers, three pooling layers (after layers 2 and 5), and two fully connected layers. This architecture is similar to the CNN-F, however, with more filters in the convolutional layers. We also use the MatConvNet implementation called *AlexNet MCN*.

(vi) The *GoogleLeNet* [52] CNN has the deepest and most complex architecture among all the other networks presented here. With two convolutional layers, two pooling layers, and nine modules also called "inception" layers, this network was designed to avoid patch-alignment issues introducing more sparsity in the inception modules. Each module consists of six convolution layers and one pooling layer concatenating these filters of different sizes and dimensions into a single new filter.

In order to form the feature vector using the pretrained CNNs, all images are scaled using bicubic interpolation to the required size for each network, in the case of this work, $224 \times 224 \times 3$. The vectors obtained by the linear layers of the CNN have size of 1024×1 for the GoogleLeNet CNN and of 4096×1 for the other networks due to their architecture specificities.

2.3.3. Classical Features. To allow the CNN features comparison and evaluation, we compared them with the results obtained by some state-of-the-art feature extraction methods for the classification of colonic polyps [18] shortly explained in the next items.

(i) *BSAG-LFD*. The Blob Shape adapted Gradient using Local Fractal Dimension method combines BA-LFD features with shape and contrast histograms from the original and gradient image [45].

(ii) *Blob SC*. The Blob Shape and Contrast algorithm [44] is a method that represents the local texture structure of an image by the analyses of the contrast and shape of the segmented blobs.

(iii) *Shearlet-Weibull*. Using the Discrete Shearlet Transform this method adopts regression to investigate dependencies across different subband levels using the Weibull distribution to model the subband coefficient distribution [53].

(iv) *GWT Weibull*. The Gabor Wavelet Transform function can be dilated and rotated to get a dictionary of filters with diverse factors [18] and its frequency using different orientations is used as a feature descriptor also using the Weibull distribution.

(v) *LCVP*. In the Local Color Vector Patterns approach, a texture operator computes the similarity between neighboring pixels constructing a vector field from an image [12].

(vi) *MB-LBP*. In the Multiscale Block Local Binary Pattern approach [54], the LBP computation is done based on average values of block subregions. This approach is used for a variety image processing applications including endoscopic polyp detection and classification [12].

For the classical features, the classification accuracy is also computed using an SVM classifier, however, with the original images (without resizing) trained using the leave-one-patient-out cross-validation strategy assuring that there are no images from patients of the validation set in the training set as in [55] to make sure the classifier generalizes to unseen patients. This cross-validation is applied to the classical feature extraction methods from the literature as well

Table 2: CNN configuration for input subimages of size 227 × 227 × 3 and its respective accuracy in %.

Size of inputs	Number of convolutional filters/size								Connected layer
	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8	
227 × 227 × 3	96/11 × 11	256/5 × 5	384/3 × 3	384/3 × 3	256/3 × 3	384/3 × 3	384/3 × 3	4096/6 × 6	4096
Accuracy: 79.00									

Table 3: Accuracy results from different CNN configurations for inputs of size 128 × 128 × 3 in %.

Network index	Number of convolutional filters/size			Connected layer	Acc
	Layer 1	Layer 2	Layer 3		
CNN-01	48/7 × 7	72/4 × 4	512/5 × 5	512	76.00
CNN-02	48/11 × 11	72/5 × 5	512/6 × 6	512	84.00
CNN-03	24/11 × 11	48/5 × 5	1024/6 × 6	1024	86.00
CNN-04	24/11 × 11	72/4 × 4	2048/5 × 5	2048	80.00
CNN-05	48/11 × 11	72/5 × 5	1024/6 × 6	1024	*87.00*

Table 4: Accuracy of different strides for overlapping subimages in the CNN-05 evaluation for i-Scan1 database in %.

Stride	Number of subimages	Accuracy
1	16384	89.00
5	676	89.00
20	49	90.00
32	25	*91.00*
48	9	87.00
Random	9	87.00
Random	25	89.00
Random	49	89.00

as to the full training and off-the-shelf CNNs features. The accuracy measure is used to allow an easy comparability of results due to the high number of methods and databases to be compared.

3. Results and Discussion

3.1. CNNs Trained from Scratch. In the first experiment for the CNN full training, we first use the configuration similar to [20] that can be seen in Table 2 and it can be concluded that the accuracy result was not satisfactory (79%). This can be explained by the fact that Neural Networks involving a large number of inputs require a great amount of computation in training, requiring more data to avoid overfitting (which is not available given the size of our dataset).

For the second experiment, the hyperparameters presented in Table 3 were selected based on the works [48, 56] and empirical adjustment tests in the architecture such as changing the size and number of filters as well as the number of units in the fully connected layer were made and are also shown in Table 3. It can be seen that the architecture CNN-05 obtained the best results, therefore, chosen to perform the subsequent tests.

In the third experiment, with the CNN-05 configuration, we trained one CNN for each patient from the database (leave-one-patient-out (LOPO) cross-validation). Specifically, the results from the CNNs presented in Table 4 are the mean values of the validation set from 62 different CNNs, one for each patient, implemented using the MatConvNet framework [50]. After training the CNN, in the evaluation phase, the final decision for a 256 × 256 pixel image of the dataset is obtained by majority voting of the decisions of all 128 × 128 pixel subimages (patches). One of the advantages of this approach is the opportunity to have a set of decisions available to acquire the final decision for one image. Also, the redundancy of overlapping subimages can increase the system accuracy likewise to give the assurance of certainty for the overall decision.

As it can be seen in Table 4, first we tested with a stride of 1 extracting the maximum number of 128 × 128 subimages available, totalizing 16384 subimages for each image, resulting in an accuracy of 89.00%. This evaluation is very computationally expensive to perform, so we decided to evaluate with different strides resulting in different number of subimages as it is shown in Table 4. We also perform a random patch extraction and it can be concluded that there is not much difference between 16384 subimages or just 25 cropped subimages (accuracy of 91.00%), saving considerable computation time and achieving good results. Besides that, using the same procedure we evaluate the architecture CNN-05 for the i-Scan3 database without staining the mucosa that presented the best results among the classical features and results are presented in Table 5.

For a better comparability of results, we trained an SVM with the extracted vectors from the last fully connected layers (LFCL) and from the prior fully connected layers (PFCL) of CNN-05 as we make in the transfer learning approach explained in the next section. The vectors are extracted from 25 cropped subimages of size 128 × 128 (with stride of 32 pixels) feedforwarded into the CNN-05 subsequently used to train a support vector machine also using the LOPO cross-validation [55]. The results from this approach using the CNN-05 architecture trained with the i-Scan1 and i-Scan3 without staining the mucosa databases are presented in Table 5. As it can be seen, using the last-layer vectors to train an SVM does not improve the results, mainly

Table 5: Accuracy of CNN-05 architecture comparing to classical features for the i-Scan1 and i-Scan3 databases in %.

Methods	i-Scan1	i-Scan3
CNN-05	*91.00*	*89.00*
CNN-05 + SVM − LFCL	83.00	72.55
CNN-05 + SVM − PFCL	80.00	66.67
BSAG-LFD	86.87	82.87
Blob SC	83.33	75.22
Shearlet-Weibull	76.67	86.80
GWT-Weibull	78.67	84.28
LCVP	66.00	77.12
MB-LBP	80.67	83.37

because the amount of data is not sufficient to generate representative features to be applied into a linear classifier. However, when the CNN is fully trained, the results surpass the classical features results as can be seen also in Table 5 mostly because the last layers are more suitable to design nonlinear hyperplanes in the classification phase. However, the problem of lack of data still is an issue and using all the information in the image would be better than using cropped patches. The significance comparison between the methods will be explored in the next section. Therefore, in order to try solving this problem, we also propose the use of transfer learning by pretrained CNNs that will be also explained in the next section.

3.2. Pretrained CNNs. In this section we present the experiments made exploring the 11 different off-the-shelf CNN architectures with the classical features trying to achieve better results than the CNN trained from scratch. As well as in the CNN trained from scratch, we use the i-Scan1 without staining the mucosa database for the first experiments.

In the first experiment, we tested the use of more samples from the same image using overlapping patches by randomly cropping 25 images of size $224 \times 224 \times 3$ of each original image of size $256 \times 256 \times 3$ (resized using bicubic interpolation for the tests presented in Table 8) increasing the database from 100 to 2500 images. The obtained results after the feature extraction performed by the CNN and after the SVM training also using the LOPO cross-validation are presented in Table 6.

It can be observed that, in this case, the use of more samples from the same image does not provide any significant improvement in the results. On the average, resizing the images produces an accuracy of 87.70% while cropping the images produces an average of 84.87%. One of the explanations for this is that, in case of resized images, there is more information about the polyp to provide to the network, so the CNN can abstract more information and form a more robust and intrinsic vector from the actual features of the lesion. However, in three cases (GoogleLeNet, VGG-VD16, and AlexNet MCN), the results using smaller cropped images surpassed the results using the entire image.

In the second experiment, still using i-Scan1 without staining the mucosa database, we also tested the use of other layers of CNNs to extract features. Table 7 shows the results obtained when the vectors are extracted from the last fully connected layer and when the vectors are from the prior fully connected layer. In the case of the last layer, the results are worse (87.70% against 85.75% on average) because the vectors from the prior fully connected layer are more related to high-level features describing the natural images used for training the original CNNs that are very different from the features to describe colonic polyp images. However, in this case, the results from CNN-F and AlexNet CNN are better using the features from the last fully connected layers.

Based on the results from the two experiments explained before, we tested the methods with all the other databases using the inputs resized to size $224 \times 224 \times 3$ by bicubic interpolation and extracting the features from the prior fully connected layer. The accuracy results for the colonic polyp classification for the 8 different databases are reported in Table 8. As can be seen, the results in Table 8 are divided into three groups: off-the-shelf features, classical features, and the fusion between off-the-shelf features and classical features that will be explained as follows.

Among the 11 pretrained CNNs investigated, the CNNs that present lower performance were GoogleLeNet, CNN-S, and AlexNet MCN. These results may indicate that such networks themselves are not sufficient to be considered off-the-shelf feature extractors for the polyp classification task.

As it can be seen in Table 8, the pretrained CNN that presents the best result on average for the different imaging modalities ($\overline{X}$) is the CNN-M network trained with the MatConvNet parameters (89.74%) followed by the CNN VGG-VD16 (88.59%). These deep models with smaller filters generalize well with other datasets as it is shown in [49], including texture recognition, which can explain the better results in the colonic polyp database. However, there is a high variability in the results and thus it is difficult to draw general conclusions.

Many results obtained from the pretrained CNNs surpassed the classic feature extractors for colonic polyp classification in the literature. The database that presents the best results using off-the-shelf features is the database staining the mucosa without any i-Scan technology (¬CVC, 88.54% on average). In the case of classical features, the database with the best result on average is the database using the i-Scan3 technology without staining the mucosa (81.61%).

To investigate the differences in the results we assess the significance of them using the McNemar test [57]. By means of this test we analyze if the images from a database are classified differently or similarly when comparing two methods. With a high accuracy it is supposed that the methods will have a very similar response, so the significance level α must be small enough to differentiate between classifying an image as correct or incorrect.

The test is carried out on the databases i-Scan3 and i-Scan1 without staining the mucosa using significance level $\alpha = 0.01$ with all the off-the-shelf CNNS, all the classical features, and the CNN-05 architecture trained from scratch. The results are presented in Figure 3. It can be observed by the black squares (indicating significantly differences)

Table 6: Results from i-Scan1 database with images resized to 224 × 224 and cropped in 25 patches of size 224 × 224.

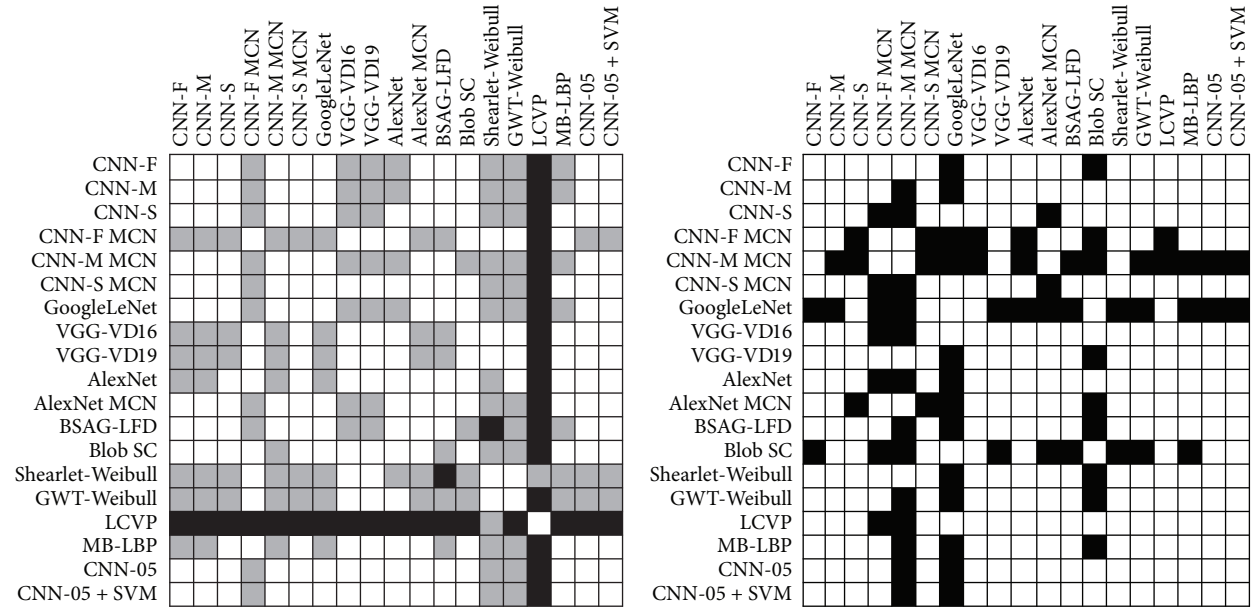
	CNN-F	CNN-M	CNN-S	CNN-FMCN	CNN-MMCN	CNN-SMCN	Google LeNet	VGG VD16	VGG VD19	AlexNet	AlexNet MCN	$\overline{X}$
Resizing image	89.33	90.67	90.00	82.00	90.67	91.42	90.67	85.33	82.67	87.33	84.67	**87.70**
Cropping 25 images	84.00	82.67	84.67	78.67	84.67	88.67	91.29	89.67	78.67	85.33	85.33	84.87

Table 7: Results from i-Scan1 database with images resized to 224 × 224 using the last fully connected layer and the prior fully connected layer.

	CNN-F	CNN-M	CNN-S	CNN-F MCN	CNN-M MCN	CNN-S MCN	Google LeNet	VGG VD16	VGG VD19	AlexNet	AlexNet MCN	$\overline{X}$
Prior fully connected layer	89.33	90.67	90.00	82.00	90.67	91.42	90.67	85.33	82.67	87.33	84.67	**87.70**
Last fully connected layer	90.67	84.67	85.33	78.67	88.00	89.33	90.67	84.67	79.33	81.33	90.67	85.75

Table 8: Accuracies of the methods for the CC-i-Scan databases in %.

Methods	No staining				Staining				
	¬CVC	i-Scan1	i-Scan2	i-Scan3	¬CVC	i-Scan1	i-Scan2	i-Scan3	$\overline{X}$
1: CNN-F	86.16	89.33	80.65	88.41	86.52	81.40	84.22	80.62	84.66
2: CNN-M	87.45	90.67	81.38	83.58	87.99	89.55	87.40	90.53	87.31
3: CNN-S	88.03	90.00	87.01	77.33	87.25	82.68	87.40	75.54	84.41
4: CNN-F MCN	88.84	82.00	73.15	90.73	85.78	89.55	89.72	83.15	85.36
5: CNN-M MCN	89.53	90.67	*88.88*	*94.66*	86.97	89.29	87.40	90.53	**89.74**
6: CNN-S MCN	90.12	*91.42*	81.38	79.85	89.18	*93.49*	81.10	84.77	86.41
7: GoogleLeNet	79.65	90.67	72.43	74.51	88.27	80.46	75.60	84.08	80.70
8: VGG-VD16	87.45	85.33	86.38	79.65	*92.47*	89.80	*95.26*	*92.38*	88.59
9: VGG-VD19	83.49	82.67	83.88	87.71	*92.47*	83.98	94.46	85.59	86.78
10: AlexNet	*91.40*	87.33	75.65	89.32	87.71	83.03	84.22	79.24	84.73
11: AlexNet MCN	89.42	84.67	78.88	83.78	89.36	83.55	81.10	78.32	83.63
$\overline{X}$	87.41	87.70	80.88	84.50	88.54	86.07	86.17	84.06	85.67
12: BSAG-LFD	*86.27*	*86.87*	*84.60*	82.87	70.20	*80.63*	78.78	71.39	**80.20**
13: Blob SC	77.67	83.33	82.10	75.22	59.28	78.83	66.13	59.83	72.79
14: Shearlet-Weibull	73.72	76.67	79.60	*86.80*	*81.30*	69.91	72.38	*83.63*	78.00
15: GWT-Weibull	79.75	78.67	70.25	84.28	*81.30*	74.54	77.17	83.39	78.66
16: LCVP	76.60	66.00	47.75	77.12	77.45	79.00	70.01	69.56	70.43
17: MB-LBP	78.26	80.67	81.38	83.37	69.29	70.60	77.22	78.32	77.38
$\overline{X}$	78.71	78.70	74.28	81.61	73.13	75.58	73.61	74.35	76.24
Fusion 5/8	88.84	85.33	83.88	92.14	93.12	90.49	96.88	94.00	90.58
Fusion 5/12	92.79	*92.67*	88.88	*96.98*	87.71	90.49	88.26	90.53	91.03
Fusion 5/8/12	*95.94*	90.00	88.88	92.14	92.30	91.43	97.63	*97.46*	93.22
Fusion 5/8/14	91.51	88.67	87.10	93.75	*94.68*	91.43	*98.44*	95.85	92.67
Fusion 5/8/15	90.91	90.00	88.88	92.14	93.94	89.80	96.88	95.61	92.27
Fusion 5/8/12/14	93.38	88.00	*91.38*	93.75	93.49	*92.12*	97.63	94.92	93.08
Fusion 5/8/12/17	93.38	90.00	*91.38*	93.75	92.75	*92.12*	97.63	*97.46*	**93.55**
CNN-05	—	91.00	—	89.00	—	—	—	—	—
CNN-05 + SVM	—	83.00	—	72.55	—	—	—	—	—



(a) i-Scan1 (b) i-Scan3

Figure 3: Results of the McNemar test for the i-Scan1 (a) and i-Scan3 (b) databases without staining. A black square in the matrix means that the methods are significantly different with significance level $\alpha = 0.01$ and a grey square in (a) means that the methods are significantly different with significance level $\alpha = 0.05$. If the square is white then there is no significant difference between the methods.

that, among the pretrained CNNs, in the i-Scan1 database the results are not significantly different and in the i-Scan3 database the CNN-M MCN and GoogleLeNet present the most significantly different results comparing to the other CNNs. It also can be seen that the CNN-05 does not have significantly different results comparing to the other CNNs in the i-Scan1 database and has significantly different results with CNN-M MCN and GoogleLeNet in the i-Scan3 database.

Also, in Figure 3, when comparing the classical feature extraction methods with the CNNs features it can be seen that there is a quite different response among the results in i-Scan3 database, especially for CNN-M MCN that is significantly different from all the classical methods with the exception of the Shearlet-Weibull method. The CNN-05 and CNN-05 + SVM did not present significantly different results with the classical features (except with LCVP in i-Scan1 database) and with the pretrained CNNs (except with CNN-M and GoogleLeNet in i-Scan3 database). Likewise, the methods with high accuracy in the i-Scan3 database (BSAG-LFD, VGG-VD16, and VGG-VD19) are not found to be significantly different.

In the i-Scan1 database, with the significance level $\alpha = 0.05$, the results are not significantly different in general (except for LCVP features). However, with the significance level $\alpha = 0.01$, the significance results represented by the grey squares in Figure 3(a) show that the two databases presented different correlation between methods which means that it is difficult to predict a good feature extractor that can satisfy both databases at the same time.

Observing the methods that presented significantly different results in Figure 3 and with good results in Table 8 we decided to produce a feature level fusion in the feature vectors concatenating them to see if the features can complement each other. It can be seen in Figure 3 that the two most successful CNNs CNN-M MCN and VGG-VD16 are significantly different from each other in both databases and the feature level fusion of these two vectors improve the results from 89.74% and 88.59%, respectively, to an accuracy of 90.58% in average as can be seen in Table 8 (Fusion 5/8).

In Figure 3(b) it can also be observed that the results from CNN-M MCN are significantly different to the classical features BSAG-LFD in the i-Scan3 database. With the feature level fusion of these two features the accuracy increases to 91.03% on average. Concatenating the three feature vectors (CNN-M MCN, VGG-VD16, and BSAG-LFD) leads to an even better accuracy: 93.22%. It is interesting to note that in both databases the results from CNN-M MCN and VGG-VD16 are significantly different. Besides that, BSAG-LFD results are significantly different to VGG-VD16 in database i-Scan1. Furthermore, BSAG-LFD results are significantly different to CNN-M MCN in database i-Scan3 which can explain the improvement in the feature level fusion between these three methods.

Making the fusion with these two off-the-shelf CNNs (CNN-M MCN and VGG-VD16) to other classical feature vectors also increases the accuracy as it can be seen in Table 8 (Fusion 5/8/14 and Fusion 5/8/15).

When we add to the vector Fusion 5/8/12 one more classical feature (MB-LBP) that is also significantly different to CNN-M MCN in database i-Scan3 and at the same time

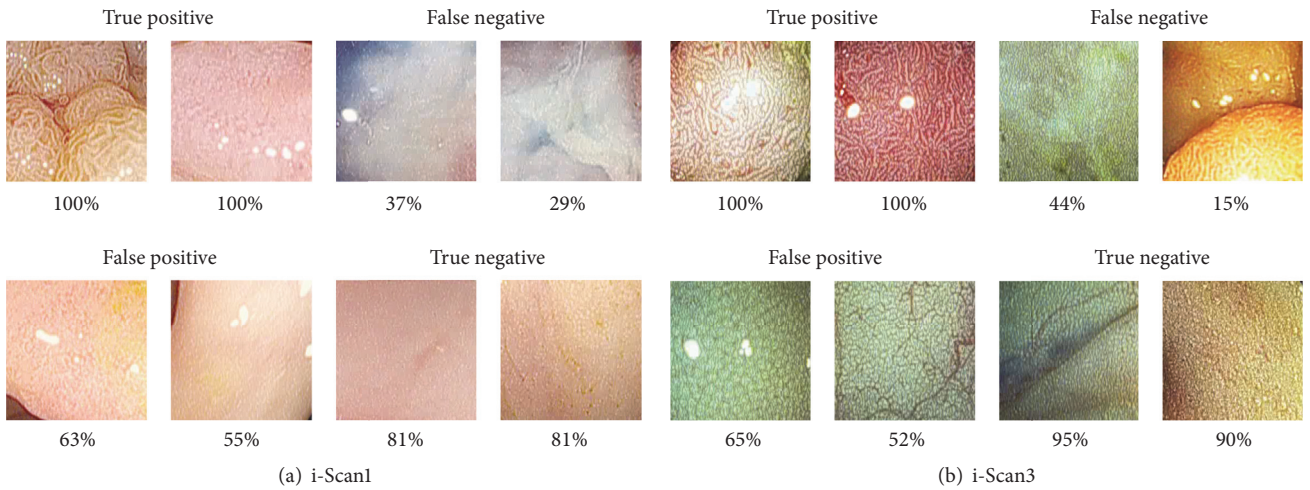


Figure 4: Example results of the classification in agreement from the methods tested in the McNemar test for each prediction outcome.

significantly different to BSAG-LFD in database i-Scan1, the result outperforms all the previous approaches: 93.55% as it can be seen in Table 8.

In Figure 4 we present some example images from the classification results of all the methods used in the McNemar test with the higher agreement for each prediction outcome. The percentage above each image shows the average classification rate of the prediction. For example, in the i-Scan1 database and i-Scan3 database (Figures 4(a) and 4(b)), the two images presented in the true positive box were classified as such in all classifiers. However, from i-Scan3 database, in the case of the false negative box, one image had 44% of misclassification and another 15% of misclassification in average.

Comparing the results from all off-the-shelf CNNs and classical features with the CNN-05 trained from scratch using the databases i-Scan1 and i-Scan3 in Table 8 it can be observed that the full training CNN outperformed the results obtained by the classical features and some of the pretrained CNNs. This approach can be considered an option for automatic colonic polyp classification, although the training time and processing complexity are not worthwhile if comparing to the off-the-shelf features.

4. Conclusion

In this work, we propose to explore Deep Learning and Transfer Learning approach using Convolutional Neural Networks (CNNs) to improve the accuracy of colonic polyp classification based on the fact that databases containing large amounts of annotated data are often limited for this type of research. For the training of CNNs from scratch, we explore data augmentation with image patches to increase the size of the training database and consequently the information to perform the Deep Learning. Different architectures were tested to evaluate the impact of the size and number of filters in the classification as well as the number of output units in the fully connected layer.

We also explored and evaluated several different pretrained CNNs architectures to extract features from colonoscopy images by knowledge transfer between natural and medical images providing what is called off-the-shelf CNNs features. We show that the off-the shelf features may be well suited for the automatic classification of colon polyps even with a limited amount of data.

Besides the fact that the pretrained CNNs were trained with natural images, the 4096 features extracted from CNN-M MCN and VGG-16 provided a good feature descriptor of colonic polyps. Some reasons for the success of the classification include the training with a large range of different images providing a powerful extractor joining the intrinsic features from the images such as color, texture, and shape in the same architecture reducing and abstracting these features in just one vector. Also, the combination of classical features with off-the-shelf features yields the best prediction results complementing each other. It can be concluded that Deep Learning using Convolutional Neural Networks is a good option for colonic polyp classification and the use of pretraining CNNs is the best choice to achieve the best results being improved by feature level fusion with classical features. In future work we plan to use this strategy to also test the detection of colonic polyps directly into video frames and evaluate the performance in real time applications as well as to use this strategy in other endoscopic databases such as automatic classification of celiac disease.

Competing Interests

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

Acknowledgments

This research was partially supported by CNPq, Brazil, for Eduardo Ribeiro under Grant no. 00736/2014-0.

References

[1] J. Bernal, J. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[2] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. de Groen, "Polyp-alert: near real-time feedback during colonoscopy," *Computer Methods and Programs in Biomedicine*, vol. 120, no. 3, pp. 164–179, 2015.

[3] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, "Texture-based polyp detection in colonoscopy," in *Bildverarbeitung für die Medizin 2009, Informatik Aktuell*, pp. 346–350, Springer, Berlin, Germany, 2009.

[4] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, and Y. A-Rahim, "A colon video analysis framework for polyp detection," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1408–1418, 2012.

[5] W. Yi, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1379–1389, 2014.

[6] N. Tajbakhsh, S. R. 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, 2016.

[7] S. Kudo, S. Hirota, T. Nakajima et al., "Colorectal tumours and pit pattern," *Journal of Clinical Pathology*, vol. 47, no. 10, pp. 880–885, 1994.

[8] M. Häfner, R. Kwitt, A. Uhl, A. Gangl, F. Wrba, and A. Vécsei, "Feature extraction from multi-directional multi-resolution image transformations for the classification of zoom-endoscopy images," *Pattern Analysis and Applications*, vol. 12, no. 4, pp. 407–413, 2009.

[9] M. Häfner, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, "Delaunay triangulation-based pit density estimation for the classification of polyps in high-magnification chromo-colonoscopy," *Computer Methods and Programs in Biomedicine*, vol. 107, no. 3, pp. 565–581, 2012.

[10] S. Kato, K. I. Fu, Y. Sano et al., "Magnifying colonoscopy as a non-biopsy technique for differential diagnosis of non-neoplastic and neoplastic lesions," *World Journal of Gastroenterology*, vol. 12, no. 9, pp. 1416–1420, 2006.

[11] S. Gross, S. Palm, J. Tischendorf, A. Behrens, C. Trautwein, and T. Aach, *Automated Classification of Colon Polyps in Endoscopic Image Data*, SPIE, Bellingham, Wash, USA, 2012.

[12] M. Häfner, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, "Color treatment in endoscopic image classification using multi-scale local color vector patterns," *Medical Image Analysis*, vol. 16, no. 1, pp. 75–86, 2012.

[13] M. Häfner, M. Liedlgruber, and A. Uhl, "Colonic polyp classification in high-definition video using complex wavelet-packets," in *Bildverarbeitung für die Medizin 2015*, Informatik Aktuell, pp. 365–370, Springer, Berlin, Germany, 2015.

[14] M. Häfner, A. Gangl, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, "Pit pattern classification using extended local binary patterns," in *Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine (ITAB '09)*, pp. 1–4, Larnaca, Cyprus, November 2009.

[15] M. Häfner, A. Uhl, A. Vécsei, G. Wimmer, and F. Wrba, "Complex wavelet transform variants and discrete cosine transform for scale invariance in magnification-endoscopy image classification," in *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB '10)*, pp. 1–5, Corfu, Greece, November 2010.

[16] Y. Yuan and M. Q.-H. Meng, "A novel feature for polyp detection in wireless capsule endoscopy images," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '14)*, pp. 5010–5015, Chicago, Ill, USA, September 2014.

[17] T. Stehle, R. Auer, S. Gros et al., "Classification of colon polyps in NBI endoscopy using vascularization features," in *Medical Imaging 2009: Computer-Aided Diagnosis*, N. Karssemeijer and M. L. Giger, Eds., vol. 7260 of *Proceedings of SPIE*, Orlando, Fla, USA, February 2009.

[18] G. Wimmer, T. Tamaki, J. J. W. Tischendorf et al., "Directional wavelet based features for colonic polyp classification," *Medical Image Analysis*, vol. 31, pp. 16–36, 2016.

[19] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '14)*, pp. 512–519, Columbus, Ohio, USA, June 2014.

[20] K. Alex, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS '12)*, pp. 1097–1105, Curran Associates, Denver, Colo, USA, 2012.

[21] H. 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.

[22] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: delving deep into convolutional nets," in *Proceedings of the 25th British Machine Vision Conference (BMVC '14)*, Nottingham, UK, September 2014.

[23] J. Guo and S. Gould, "Deep CNN ensemble with data augmentation for object detection," https://arxiv.org/abs/1506.07224.

[24] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 1717–1724, Columbus, Ohio, USA, June 2014.

[25] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," in *Proceedings of the IEEE 12th International Symposium on Biomedical Imaging (ISBI '15)*, pp. 294–297, April 2015.

[26] B. Van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *Proceedings of the 12th IEEE International Symposium on Biomedical Imaging (ISBI '15)*, pp. 286–289, Brooklyn, NY, USA, April 2015.

[27] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Proceedings of the 12th IEEE International Symposium on Biomedical Imaging (ISBI '15)*, pp. 79–83, New York, NY, USA, April 2015.

[28] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013*, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds., vol. 8150 of *Lecture Notes in Computer Science*, pp. 411–418, Springer, Berlin, Germany, 2013.

[29] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 2843–2851, December 2012.

[30] N. Tajbakhsh, M. B. Gotway, and J. Liang, "Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part II*, vol. 9350 of *Lecture Notes in Computer Science*, pp. 62–69, Springer, Berlin, Germany, 2015.

[31] H. R. Roth, L. Lu, A. Seff et al., *A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations*, Springer International, Cham, Switzerland, 2014.

[32] R. Zhu, R. Zhang, and D. Xue, "Lesion detection of endoscopy images based on convolutional neural network features," in *Proceedings of the 8th International Congress on Image and Signal Processing (CISP '15)*, pp. 372–376, Shenyang, China, October 2015.

[33] H. Roth, J. Yao, L. Lu, J. Stieger, J. Burns, and R. Summers, Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications, CoRR, abs/1407.5976, 2014.

[34] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu et al., "Convolutional neural networks for medical image analysis: full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[35] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "A comprehensive computer-aided polyp detection system for colonoscopy videos," in *Proceedings of the 24th International Conference on Information Processing in Medical Imaging (IPMI '15)*, pp. 327–338, Sabhal Mor Ostaig, Isle of Skye, UK, June-July 2015.

[36] N. Hatipoglu and G. Bilgin, "Classification of histopathological images using convolutional neural network," in *Proceedings of the 4th International Conference on Image Processing Theory, Tools and Applications (IPTA '14)*, pp. 1–6, October 2014.

[37] Y. Zou, L. Li, Y. Wang, J. Yu, Y. Li, and W. J. Deng, "Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network," in *Proceedings of the IEEE International Conference on Digital Signal Processing (DSP '15)*, pp. 1274–1278, IEEE, Singapore, July 2015.

[38] J. S. Yu, J. Chen, Z. Q. Xiang, and Y. X. Zou, "A hybrid convolutional neural networks with extreme learning machine for WCE image classification," in *Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO '15)*, pp. 1822–1827, IEEE, Zhuhai, China, December 2015.

[39] E. Ribeiro, A. Uhl, and M. Häfner, "Colonic polyp classification with convolutional neural networks," in *Proceedings of the IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS '16)*, pp. 253–258, Dublin, Ireland, June 2016.

[40] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2013—16th International Conference, Nagoya, Japan, September 2013, Proceedings, Part II*, pp. 246–253, Springer, 2013.

[41] F. Ciompi, B. de Hoop, S. J. van Riel et al., "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box," *Medical Image Analysis*, vol. 26, no. 1, pp. 195–202, 2015.

[42] J. Arevalo, F. A. Gonzalez, R. Ramos-Pollan, J. L. Oliveira, and M. A. Guevara Lopez, "Convolutional neural networks for mammography mass lesion classification," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '15)*, pp. 797–800, IEEE, Milan, Italy, August 2015.

[43] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[44] M. Häfner, A. Uhl, and G. Wimmer, "A novel shape feature descriptor for the classification of polyps in HD colonoscopy," in *Medical Computer Vision. Large Data in Medical Imaging*, B. Menze, G. Langs, A. Montillo, M. Kelm, H. Müller, and Z. Tu, Eds., vol. 8331 of *Lecture Notes in Computer Science*, pp. 205–213, Springer, Berlin, Germany, 2014.

[45] M. Häfner, T. Tamaki, S. Tanaka, A. Uhl, G. Wimmer, and S. Yoshida, "Local fractal dimension based approaches for colonic polyp classification," *Medical Image Analysis*, vol. 26, no. 1, pp. 92–107, 2015.

[46] H. R. Roth, L. Lu, J. Liu et al., "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1170–1181, 2015.

[47] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2144–2151, 2012.

[48] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the 4th International Conference on Artificial Intelligence and Statistics (AISTATS '11)*, vol. 15 of *JMLR*, pp. 215–223, JMLR W&CP, Fort Lauderdale, Fla, USA, 2011.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, abs/1409.1556, 2014.

[50] A. Vedaldi and K. Lenc, "Matconvnet—convolutional neural networks for MATLAB," CoRR, abs/1412.4564, 2014.

[51] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," CoRR, abs/1312.6229, 2013.

[52] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 1–9, Boston, Mass, USA, June 2015.

[53] Y. Dong, D. Tao, X. Li, J. Ma, and J. Pu, "Texture classification and retrieval using shearlets and linear regression," *IEEE Transactions on Cybernetics*, vol. 45, no. 3, pp. 358–369, 2015.

[54] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Li, "Learning multi-scale block local binary patterns for face recognition," in *Advances in Biometrics*, vol. 4642 of *Lecture Notes in Computer Science*, pp. 828–837, Springer, Berlin, Germany, 2007.

[55] M. Häfner, M. Liedlgruber, S. Maimone, A. Uhl, A. Vécsei, and F. Wrba, "Evaluation of cross-validation protocols for the classification of endoscopic images of colonic polyps," in *Proceedings of the 25th IEEE International Symposium on*

Computer-Based Medical Systems (CBMS '12), pp. 1–6, Rome, Italy, June 2012.

[56] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Intelligent Signal Processing*, pp. 306–351, IEEE Press, Piscataway, NJ, USA, 2001.

[57] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 4345936, 10 pages
http://dx.doi.org/10.1155/2016/4345936

Research Article

An Active Learning Classifier for Further Reducing Diabetic Retinopathy Screening System Cost

Yinan Zhang[1,2] and Mingqiang An[3]

[1]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*
[2]*College of Computer Science and Information Engineering, Tianjin University of Science and Technology, Tianjin 300222, China*
[3]*College of Science, Tianjin University of Science and Technology, Tianjin 300222, China*

Correspondence should be addressed to Yinan Zhang; zh.y.n@163.com

Received 7 May 2016; Revised 24 June 2016; Accepted 26 July 2016

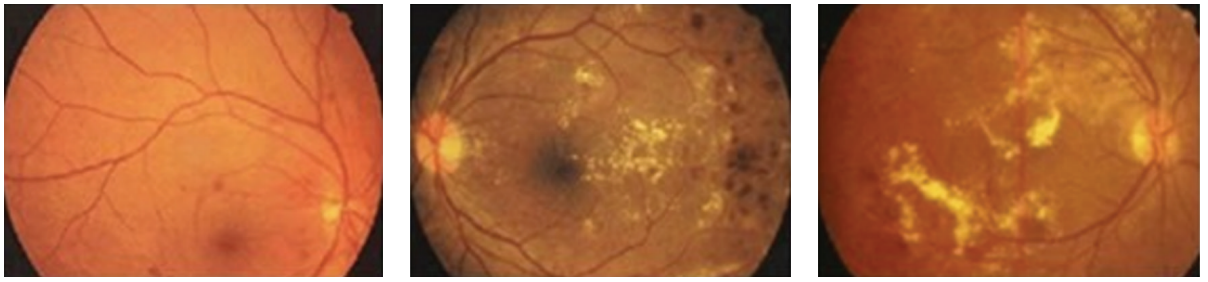
Academic Editor: Georgy Gimel'farb

Diabetic retinopathy (DR) screening system raises a financial problem. For further reducing DR screening cost, an active learning classifier is proposed in this paper. Our approach identifies retinal images based on features extracted by anatomical part recognition and lesion detection algorithms. Kernel extreme learning machine (KELM) is a rapid classifier for solving classification problems in high dimensional space. Both active learning and ensemble technique elevate performance of KELM when using small training dataset. The committee only proposes necessary manual work to doctor for saving cost. On the publicly available Messidor database, our classifier is trained with 20%–35% of labeled retinal images and comparative classifiers are trained with 80% of labeled retinal images. Results show that our classifier can achieve better classification accuracy than Classification and Regression Tree, radial basis function SVM, Multilayer Perceptron SVM, Linear SVM, and K Nearest Neighbor. Empirical experiments suggest that our active learning classifier is efficient for further reducing DR screening cost.

1. Introduction

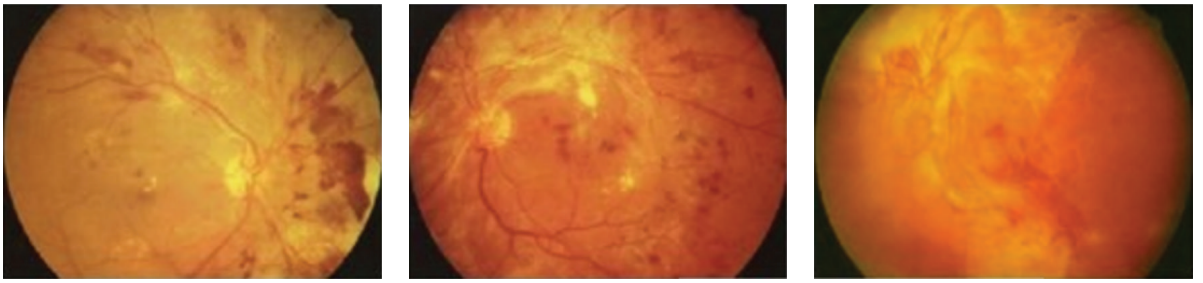
Diabetic retinopathy (DR) [1] is one of the most common causes of blindness in diabetic mellitus research [2]. Millions of diabetic patients suffer from DR. DR not only deprives patients' sight [3] but also brings heavy burden to their family and society [4]. In 2012 [5], 29.1 million Americans (9.3% of the population) were diagnosed with diabetes. A more serious problem is that 76% of those patients were becoming with worsening diabetes. Each year, approximately 1.4 million Americans are diagnosed with diabetes. With the development of diabetes, about 40% of patients may lose sight from DR [6]. Recently, new technique named optical coherence tomography (OCT) is popular in developed countries. OCT can perform cross-sectional imaging, but OCT is still too expensive for many areas which are economically underdeveloped. Thus DR screening system is still useful for diabetic patients in many low income areas. This challenging problem causes a demand of a better computer-aided DR screening system [7, 8].

Many computer-aided screening systems can reduce massive manual screening effectively [9, 10]. Gardner et al. [11] propose an automatic DR screening system with artificial neural network. Most of computer-aided DR screening researches focus on reducing and improving doctor's work. It is noteworthy that Liew et al. [12] point out a critical issue; this issue is about accuracy and cost effectiveness. A typical DR screening hardware system includes but is not limited to high resolution camera, computing system, and storage system. The software system for DR screening system mainly contains three major parts: image processing [13], feature extraction [14], and classification [15] (automatic diagnosis result of computer). The architecture of computer-aided DR screening hardware system is clear and stable nowadays, but software system still has much space for development. Classification is an important breakthrough for improving DR screening system, especially when applying active learning method rather than supervised learning or unsupervised learning method.



(a) DR grade I (b) DR grade II (c) DR grade III

Figure 1: Representative images having different grades (I, II, and III).



(a) DR grade IV (b) DR grade V (c) DR grade VI

Figure 2: Representative images having different grades (IV, V, and VI).

However, to build an automatic computer-aided screening system raised a financial problem [16]. A DR screening system faces three major requirements nowadays. First, when a company builds a DR screening system for medical purpose, the accuracy is a key measurement. Second, hospital administrators need that this DR system not only can make classification automatically but also can save more money and time when it is running in the future. Third, the DR screening system should raise meaningful queries to doctors as many as possible, and cases that can be easily diagnosed by computer should be queried as little as possible. Therefore, a DR screening system should further have the following three characters: (1) more accuracy, (2) smaller training dataset, and (3) active learning.

For solving the above problems, we propose an ensemble-kernel extreme learning machine (KELM) based active learning with querying by committee classifier. Below are the major contributions/conclusions of our work:

(1) Retinal image is easy to snap, but manually diagnosing a result is of high cost.

(2) Kernel technique is suitable for classifying retinal images which is related to classification in high dimensional spaces.

(3) Ensemble learning (bagging technique) can elevate classifier's performance. Particularly, overfitting occurs when training set is small .

(4) Active learning can further reduce the size of training dataset compared to traditional machine learning method in DR screening system.

(5) The committee can avoid unnecessary queries to doctor; this is distinctive to other state-of-the-art DR screening systems.

This paper is organized as follows: Section 2 shows background of retinal images and related works, Section 3 presents the details of the proposed classifier, and Section 4 presents empirical experiment and results. Conclusions are drawn in the final section.

2. Retinal Images and Related Works

2.1. Retinal Image and Detections. Figure 1 shows DR grade [17]: I, II, and III. Figure 2 shows DR grade: IV, V, and VI. Microaneurysm appears as tiny red dots in Figure 1; with the worsening of diabetes, exudates occur as primary signs of diabetic retinopathy. In Figures 1 and 2, inhomogeneity appears and it can lead to loss of sight.

Doctors give diagnosis results based on 3 major lesions: microaneurysm, exudates, and inhomogeneity. Moreover, there are two useful anatomical detections: macula and optic disc. In Table 1, five essential detections of DR screening are listed.

2.2. Classic DR Screening System Architecture. DR screening system [19] captures retinal images and gives diagnosis results. The classic architecture of DR screening system is shown in Figure 3.

A high resolution camera is used for capturing retinal images. Then, retinal images are saved into storage system. Usually, there is a preprocessing for retinal images; this

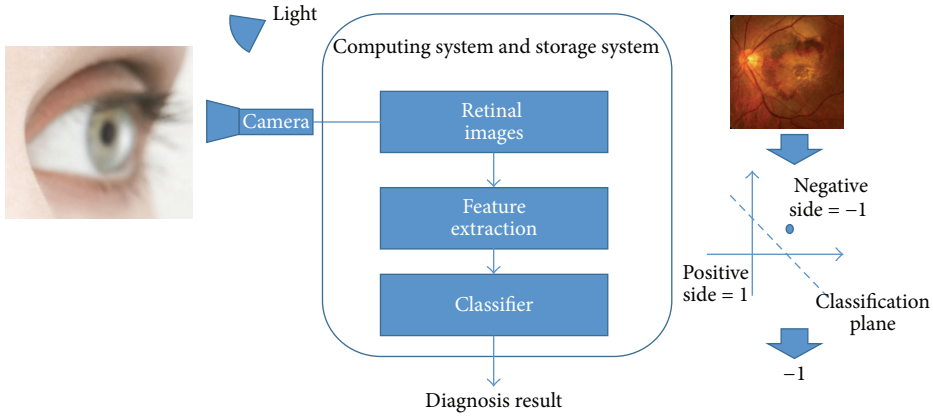


Figure 3: DR screening system architecture.

process enhances image contrast and so forth. In the next step, multiple features of retinal images are extracted by image algorithms. Extracted features are represented in high dimensional space. Therefore original retinal image is mapped into high dimensional space. One retinal image is presented as a vector (or a dot) in this high dimensional space. Finally, a trained classifier gives a binary result (−1/1 or 1/0). This binary result indicates that the vector belongs to the "positive" side or the "negative" side. Figure 3 also exemplifies a brief workflow in two-dimensional space.

Many DR screening system studies focus on the performance of accuracy measurement. Dabramoff et al. [20] pointed out that DR screening system is an investigated field. Fleming et al. [21] showed that reducing mass manual effort is the key of creating DR screening system. Meanwhile, several researchers focus on automatic diagnosis of patients having DR [22]. Even though those researches and applications save massive manual work, DR screening system cost can be further reduced.

Table 1: Essential detection of DR screening.

Detection target	Detail information
Microaneurysm	An extremely small aneurysm, it looks as tiny red dots in retinal image.
Exudates	Fat or lipid leak from aneurysms or blood vessels, it looks as small and bright spots with irregular shape.
Inhomogeneity	Regions of retina are different and unusual.
Macula	The macula is an oval-shaped pigmented area near the center of the retina of the human eye.
Optic disc	The optic disc is the point of exit for ganglion cell axons leaving the eye.

3. Ensemble Extreme Learning Machine Based Active Learning Classifier with Query by Committee

In this section, the proposed classifier is described in detail. With the consideration of accuracy, time consuming, computing resource consuming, high dimensional features classification, and reducing artificial labeling, we adapt kernel extreme learning machine (KELM) and then we use ensemble learning (bagging technique) to solve overfitting problem. Moreover, the bagging-KELM can be trained in parallel computing architecture.

3.1. Active Learning with Query by Committee. Active learning [23] has control over instances, once active learning reaches query paradigm, in which the committee can assign new artificial labeling task for human. Query by committee (QBC) [24] is a learning method which adopts decision of a committee to decide an unlabeled instance should be asked for artificial labeling or not. Once an artificial labeling task is finished, the new artificial labeled instance is added into training set. Therefore, the committee reduces testing instances and enlarges training set with asking for artificial labeling work. Since QBC has control over instances from which it learns, QBC maintains a group of hypotheses from training set; those hypotheses represent the version space. For real word problems, the size of committee should be big enough.

Figure 4 shows the proposed method. Our approach contains 3 cyclic steps.

Step 1. KELM with bagging technique and committee are trained synchronously. Initial training instances consist of extracted features from DR images and corresponding artificial label marks.

Step 2. After the training procedure, the committee can propose necessary queries for bagging-KELM.

Step 3. In the testing procedure, both bagging-KELM and the committee receive testing instances and then bagging-KELM asks permission from the committee. If committee agrees with bagging-KELM, bagging-KELM gives a hypothesis for an unlabeled instance as final diagnosis result. However, if committee gives disagreement, the committee proposes

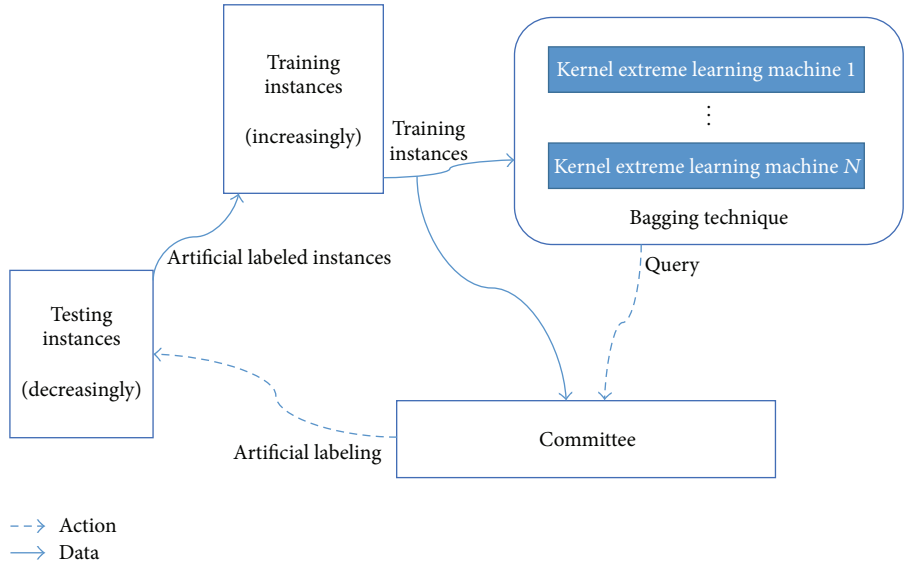


Figure 4: Active learning with query with committee.

the unlabeled retinal image to human doctor (this increases training instances).

To conclude, our approach is dealing with 3 optimization problems: (1) increasing training dataset as little as possible, (2) increasing training dataset with necessary queries, and (3) decreasing testing dataset with control.

3.2. Kernel Extreme Learning Machine. Extreme learning machine (ELM) [25] is a fast and accurate single-forward layer feedforward neural network classification algorithm proposed by Huang et al. Different from traditional neural networks, ELM assigns perceptron with random weights in the input layer and then the weights of output layer can be calculated catalytically by finding the least square solution. Therefore, ELM is faster than other learning algorithms for neural network; the time cost is extremely low.

For diabetic retinopathy screening, given a training dataset X with N labeled instances (x_i, t_i), $i = 1, 2, \ldots, N$, where each x_i is an n dimensional vector, $x_i = [x^1, x^2, x^3, \ldots, x^n]^T \in R^n$, and t_i is an indicating label of corresponding instance, the output of signal-layer forward network with M perceptrons in middle layer can be calculated as follows:

$$\sum_{i=1}^{M} \beta_i g_i\left(x_j\right) = \sum_{i=1}^{M} \beta_i g\left(w_i + x_j + b_i\right) = t_j, \quad j = 1, 2, \ldots, N, \tag{1}$$

where w_i is the weights connecting the ith middle perceptron with the input perceptron. β_i is the weights connecting the ith hidden perceptron with the output perceptron, and b_i is the bias of the ith hidden perceptron.

$g(\cdot)$ denotes nonlinear activation function; some classical activation functions are listed as follows:

(1) Sigmoid function:

$$G(a, b, x) = \frac{1}{1 + \exp\left(-\left(a \cdot x + b\right)\right)}. \tag{2}$$

(2) Fourier function:

$$G(a, b, x) = \sin\left(a \cdot x + b\right). \tag{3}$$

(3) Hard limit function:

$$G(a, b, x) = \begin{cases} 1, & \text{if } a \cdot x - b \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

(4) Gaussian function:

$$G(a, b, x) = \exp\left(-b \left\|x - a\right\|^2\right). \tag{5}$$

(5) Multiquadrics function:

$$G(a, b, x) = \left(\left\|x - a\right\|^2 + b^2\right)^{1/2}. \tag{6}$$

Equation (1) can be expressed in a compact equation as follows:

$$H\beta = T, \tag{7}$$

where H is the middle layer output matrix:

$$H = \begin{bmatrix} g\left(w_1 x_1 + b_1\right) & \cdots & g\left(w_N x_1 + b_N\right) \\ \vdots & \ddots & \vdots \\ g\left(w_1 x_M + b_1\right) & \cdots & g\left(w_N x_M + b_N\right) \end{bmatrix}, \tag{8}$$

where β is the matrix of middle-to-output weights and T is the target matrix.

In (8), weights w_i and bias b_i are assigned random float number and $g(\cdot)$ is selected as sigmoid function; therefore the output of middle perceptron can be determined very fast, which is H in (7).

The remaining work is minimum square error estimation:

$$\min_{\beta} \|H\beta - T\|. \tag{9}$$

The smallest norm least squares solution for (9) can be calculated by applying the definition of the Moore-Penrose generalized inverse; the solution is as follows:

$$\widehat{\beta} = H^{-1}T, \tag{10}$$

where H^{-1} is the generalized inverse of matrix H.

The least squares solution of (10) based on Kuhn-Tucker conditions can be written as follows:

$$\beta = H^T \left(\frac{1}{C} + HH^T\right)^{-1} T, \tag{11}$$

where H is the middle layer output, C is regulation coefficient, and T is the expected output matrix of instances.

Therefore, the output function is

$$f(x) = h(x) H^T \left(\frac{1}{C} + HH^T\right)^{-1} T. \tag{12}$$

The kernel matrix of ELM can be defined as follows:

$$M = HH^T : m_{ij} = h(x_i) h(x_j) = k(x_i, x_j). \tag{13}$$

Therefore, the output function $f(x)$ of kernel extreme learning machine can be expressed as follows:

$$f(x) = [k(x, x_1), \ldots, k(x, x_N)] \left(\frac{1}{C} + M\right)^{-1} T, \tag{14}$$

where $M = HH^T$ and $k(x, y)$ is the kernel function of perceptrons in middle layer.

We adopt three kernel functions in this paper; they are as follows:

POLY: for some positive integer d,

$$k(x, y) = (1 + \langle x, y\rangle)^d. \tag{15}$$

RBF: for some positive number a,

$$k(x, y) = \exp\left(-\frac{\langle (x - y), (x - y)\rangle}{(2a2)}\right). \tag{16}$$

MLP: for a positive number p and negative number q,

$$k(x, y) = \tanh(p \langle x, y\rangle + q). \tag{17}$$

Compared with ELM, KELM performs similarly to or better than ELM, and KELM is more stable [25]. Compared with SVM, KELM spends much less time without performance loses.

3.3. Bagging Technique. By applying ensemble learning [26] to our approach, classifier can obtain better classification performance when dealing with overfitting problem brought by small training set. We apply bagging technique to enhance KELM classifier. Bagging technique seeks to promote diversity among the methods it combines. In the initialization procedure, we adopt multiple different kernel functions and different parameters. Therefore, a group of classifiers can be built for bagging technique implementation.

When applying a group of KELMs with bagging method, each KELM is trained independently and then those KELMs are aggregated via a majority voting technique. Given a training set TR $= \{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$, where x_i is extracted features from retinal images and y_i is corresponding diagnosis result, we then build M training datasets randomly to construct M KELMs bagging independently.

The bootstrap technique is as follows:

init:

given training datasets TR

M distinctive KELM$_i$, $i = 1, 2, 3, \ldots, M$

training:

construct subtraining datasets {STR$_i$ | $i = 1, 2, \ldots, M$} form TR with resampling randomly and replacement

train KELM$_i$ with STR$_i$

classification:

calculate hypothesis H_i of ELM$_i$, $i = 1, 2, 3, \ldots, M$

perform majority voting of H_i

4. Experiments and Results

4.1. Messidor Database and Evaluation Criteria. For empirical experiment, we use public Messidor dataset [27] that consists of 1151 instances. Images are of 45-degree field of view and three different resolutions ($440 * 960$, $2240 * 1488$, and $2304 * 1536$).

Each image is labeled 0 or 1 (negative or positive diagnostic result). 540 images are labeled 0; the remnants are labeled 1. Many researches did 5-fold (or 10-fold) cross-validation. Thus, 80% of database is training dataset and the remaining 20% instances are testing dataset. We train Classification and Regression Tree (CART), radial basis function (RBF) SVM, Multilayer Perceptron (MLP) SVM, Linear (Lin) SVM, and K Nearest Neighbor (KNN) with 80% of database and the remaining 20% is as testing dataset.

For the proposed active learning (AL) classifier, we use 10%–20% of database as initial training dataset and give it 10%–15% of database as queries made by committee. Therefore, 20%–35% of database is used to train active learning classifier in total, and the remaining 65%–80% of database is testing dataset. We also train ELM and KELM with 20%–35% of database, and the remaining 65%–80% of database is

testing dataset. Therefore, ELM, KELM, and our approach are trained with the same amount of labeled instances; the results can prove the availability of kernel technique, bagging technique, and active learning. Committee contains all classifiers which were mentioned in this paper.

In short, we use 80% of Messidor database to train 5 classifiers, and we cut more than half the training instances to validate ELM, KELM, and our approach. Details are presented in Section 4.3. Each classifier was tested 10 times. The recommendations of the British Diabetic Association (BDA) are 80% sensitivity and 95% specificity. Therefore, accuracy, sensitivity, and specificity are compared among those classifiers.

Sensitivity, accuracy, and specificity are defined as follows:

$$\begin{aligned} \text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \\ \text{Specificity} &= \frac{\text{TN}}{\text{FP} + \text{TN}}, \end{aligned} \tag{18}$$

where TP, FP, TN, and FN are the true and false positive and true and false negative classifications of a classifier.

4.2. Retinal Image Features. Table 2 lists every feature extracted from retinal images with feature information. The retinal images are mapped in a 19-dimensional space.

The details of image features used in Messidor database are listed as follows:

(1) *Quality Assessment*. Messidor database contains sufficient quality image for a reliable diagnosis result. After detecting vessel system, the box count values can be calculated for a supervised learning classifier. The vessel segmentation algorithm is based on Hidden Markov Random Fields (HMRF) [28].

(2) *Prescreening*. Images are classified as abnormal or to be needed for further processing. Every image is split into disjoint subregions and inhomogeneity measure [29] is extracted for each subregion. Then a classifier learns from these features and classifies the images.

(3) *MA Detection*. Microaneurysms appear as small red dots and they are hard to find efficiently. The MA detection method used in Messidor database is based on preprocessing method and candidate extractor ensembles [30].

(4) *Exudate*. Exudates are bright small dots with irregular shape. By following the likely complex methodology as for microaneurysm detection [30], it combines preprocessing methods and candidate extractors for exudate detection [31].

(5) *Macula Detection*. Macula is located in the center of the retina. By extracting the largest object from image with brighter surroundings [32], the macula can be detected effectively.

Table 2: Image features of Messidor dataset [18].

Feature	Feature information
(0)	The binary result of quality assessment. 0: bad quality; 1: sufficient quality.
(1)	The binary result of prescreening, where 1 indicates severe retinal abnormality and 0 its lack.
(2–7)	The results of MA detection. Each feature value stands for the number of MAs found at the confidence levels alpha = $0.5 \cdots 1$, respectively.
(8–15)	Contain the same information as (2–7) for exudates. However, as exudates are represented by a set of points rather than the number of pixels constructing the lesions, these features are normalized by dividing the number of lesions with the diameter of the ROI to compensate different image sizes.
(16)	The Euclidean distance of the center of the macula and the center of the optic disc to provide important information regarding the patient's condition. This feature is also normalized with the diameter of the ROI.
(17)	The diameter of the optic disc.
(18)	The binary result of the AM/FM-based classification.
(19)	Class label. 1: containing signs of DR (accumulative label for the Messidor classes 1, 2, and 3); 0: no signs of DR.

(6) *Optic Disc Detection*. Optic disc is anatomical structure with circular shape. Ensemble-based system of Qureshi et al. [33] is used for optic disc detection.

(7) *AM/FM-Based Classification*. The Amplitude-Modulation Frequency-Modulation method [34] decomposes the green channel of the image and then signal processing techniques are applied to obtain representations which reflect the texture, geometry, and intensity of the structures.

4.3. Experiment Results. CART, RBF, MLP, Lin, KNN, AL, KELM, and ELM are compared in experiments. For CART, RBF, MLP, Lin, and KNN, 80% of labeled retinal images are offered to these 5 classifiers as training dataset. The parameters of those 5 classifiers are determined by grid search on training dataset. To AL, KELM and ELM are also used as grid-search method to set hidden layer. The MATLAB R2015a version is used in this paper.

Figure 5 shows the boxplot of normalized correct classifications. In Figure 5(a), AL has 115 instances (10% of Messidor database) for initial training and 115 more instances (10% of Messidor database) are queries from committee. Therefore, 230 (20% of Messidor database) instances are used in training AL in total. For testing kernel, bagging technique, and active learning, KELM and ELM are offered 230 training instances (20% of Messidor database). Other 5 classifiers are trained with 920 instances (80% of Messidor database).

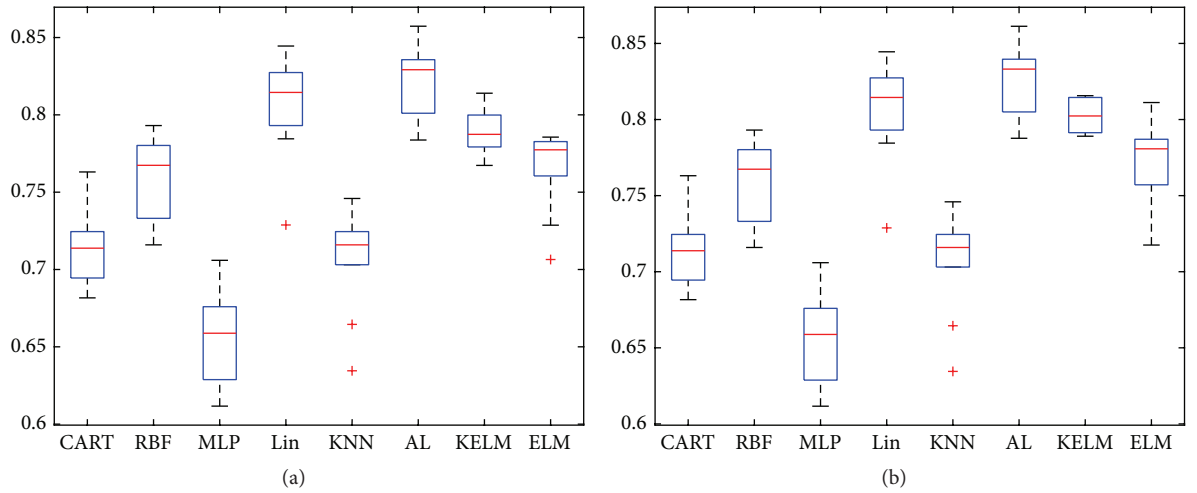


Figure 5: (a) Using 10% of Messidor dataset as initial training dataset and the committee proposes 10% of dataset as queries and (b) using 10% of Messidor dataset as initial training dataset and the committee proposes 15% of dataset as queries.

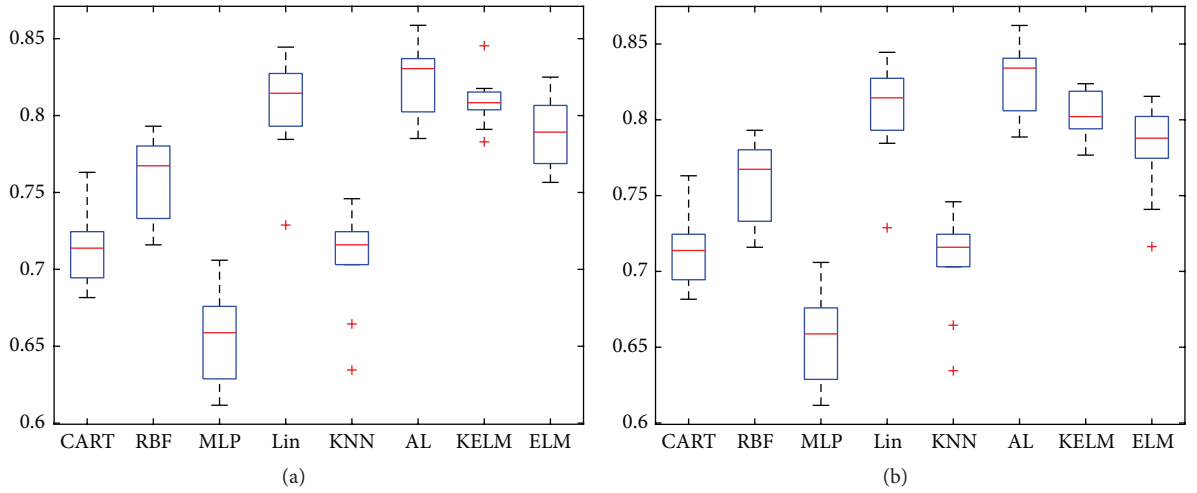


Figure 6: (a) Using 15% of Messidor dataset as initial training dataset and the committee proposes 10% of dataset as queries and (b) using 15% of Messidor dataset as initial training dataset and the committee proposes 15% of dataset as queries.

KELM is classified more accurately than ELM by the kernel technique. Bagging technique and active learning method further elevate classification accuracy of KELM. Comparing AL with other 7 classifiers, its correct classification is about 2%~20% higher than other classifiers in Figure 5(a), and the training dataset of AL is only 25% of other 5 classifiers. MLP, CART, and KNN are the worst three classifiers. RBF performs a little better than ELM, but RBF has three times more labeled instances than ELM. Lin performs better than ELM and KELM, but it is slightly lower than AL.

Similarly, in Figure 5(b), active learning and other classifiers have been tested again. KELM gives more correct classification results and AL is better than both ELM and KELM. Comparing AL with other classifiers, AL achieves better classification accuracy and AL only needs 287 labeled instances for training.

In Figure 6, CART, RBF, MLP, Lin, and KNN are exactly the same as in Figure 5. In Figure 6(a), a bigger initial training dataset (15%) is used to train AL and 25% labeled instances are given to KELM and ELM as training dataset. In Figure 6(b), 30% labeled instances are given to KELM and ELM for training. In Figure 6, kernel technique helps ELM to produce more correct classification results, and the active learning method still further boosts KELM. In Figure 6, Lin performs closely to AL, but Lin has nearly triple the size of training dataset. Therefore, the disadvantage of Lin is the need of massive manual work.

Figure 7 shows 20% of labeled instances as initial training dataset for AL. Figure 7 proves the same conclusion as shown in Figures 5-6; kernel technique, bagging technique, and active learning are effective and efficient for improving ELM. It should be noticed that KELM and ELM are tested twice

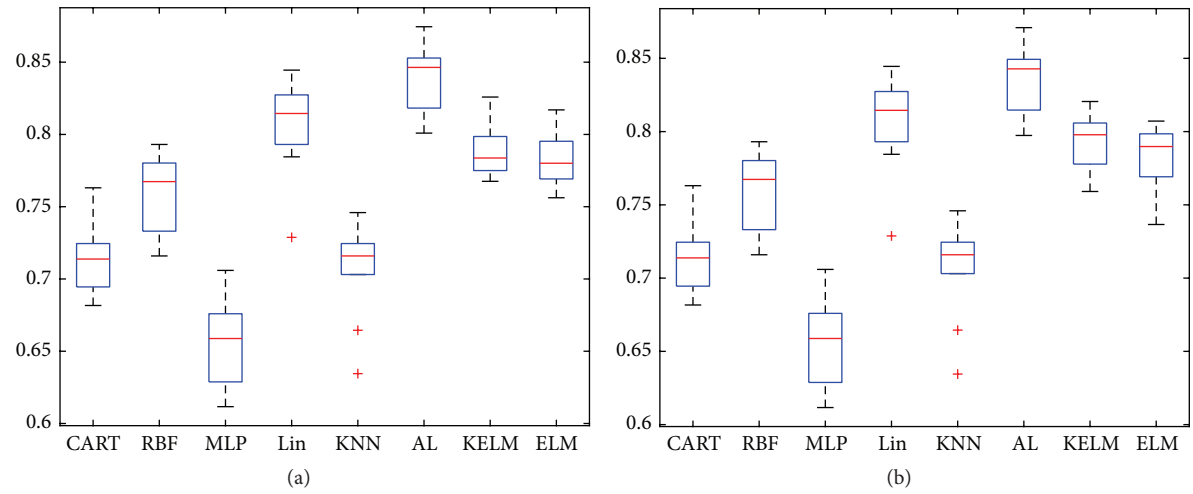


Figure 7: (a) Using 20% of Messidor dataset as initial training dataset and the committee proposes 10% of dataset as queries and (b) using 20% of Messidor dataset as initial training dataset and the committee proposes 15% of dataset as queries.

Table 3: Details of Figures 5–7 (accuracy).

	Classifiers	Max	Min	Mean
Figures 5–7	CART_80	0.775	0.693	0.725
	RBF_80	0.805	0.728	0.770
	MLP_80	0.718	0.623	0.669
	Lin_80	0.856	0.741	0.818
	KNN_80	0.758	0.646	0.720
Figure 5(a)	AL_10_10	0.872	0.799	0.838
	KELM_20	0.832	0.785	0.808
	ELM_20	0.804	0.725	0.783
Figure 5(b)	AL_10_15	0.880	0.807	0.846
	KELM_25	0.803	0.776	0.790
	ELM_25	0.799	0.705	0.761
Figure 6(a)	AL_15_10	**0.886**	0.804	0.836
	KELM_25	0.834	0.771	0.798
	ELM_25	0.813	0.745	0.778
Figure 6(b)	AL_15_15	0.871	0.797	**0.851**
	KELM_30	0.812	0.765	0.792
	ELM_30	0.804	0.705	0.769
Figure 7(a)	AL_20_10	0.874	0.800	0.839
	KELM_30	0.833	0.774	0.795
	ELM_30	0.824	0.763	0.790
Figure 7(b)	AL_20_15	0.878	**0.812**	0.843
	KELM_35	0.837	0.776	0.811
	ELM_35	0.823	0.753	0.800
Max of column		**0.886**	**0.812**	**0.851**

Table 4: Sensitivity and specificity.

Classifiers	Sensitivity mean	Specificity mean
CART_80	77.64%	83.10%
RBF_80	78.07%	86.17%
MLP_80	74.41%	84.52%
Lin_80	80.29%	88.96%
KNN_80	77.13%	88.13%
AL_10_10	81.69%	91.46%
KELM_20	79.44%	90.81%
ELM_20	78.92%	90.03%
AL_10_15	82.38%	91.54%
KELM_25	78.43%	90.72%
ELM_25	77.80%	90.26%
AL_15_10	82.67%	92.11%
KELM_25	80.54%	91.23%
ELM_25	79.87%	90.78%
AL_15_15	82.63%	92.08%
KELM_30	81.88%	91.91%
ELM_30	81.35%	90.35%
AL_20_10	82.78%	91.58%
KELM_30	81.83%	90.03%
ELM_30	81.10%	89.61%
AL_20_15	82.63%	91.72%
KELM_35	81.95%	90.18%
ELM_35	81.21%	88.96%

which Figures 5(b) and 6(a) present. Similarly, Figures 6(b) and 7(a) also present twice the comparison results of ELM and KELM. Table 4 lists results of Figures 5–7 in detail.

Table 3 contains 5 columns; the names of classifiers are attached with experiment parameters. For instance, KNN_80 is that KNN classifier is trained with 80% of instances. The max, min, and mean are calculated from 10 runs. AL_10_15 is that 10% of labeled instances are as initial training dataset and 15% of labeled instances are queries form committee.

In Table 3, the lower limit and mean value of AL_10_10 are the highest in column. The upper limit of AL_20_10 is the highest in column.

In Table 4, mean values of sensitivity and specificity are listed for all classifiers. The first column of Table 4 is

corresponding experiment as Table 3. Second column is mean values of sensitivity, and third column is mean values of specificity. All mean values are statistical result of 10 runs. Sensitivity mean values are between 0.74 and 0.82; specificity mean values are between 0.83 and 0.92.

4.4. Discussions. In this section, we present two issues about experiment results: (1) what are the advantages of KELM? (2) Is the proposed method suitable for medical implement?

The KELM is ELM with kernel technique; this approach is similar to SVM. The kernel technique can map original data (linear inseparable) into a new space (higher dimensional space but linear separable) for a linear classifier. The major contribution of KELM is that kernel technique helps ELM to face a high dimensional classification problem which is faster than kernel-SVM when solving the same problem. Especially in this paper, the Messidor dataset contains 18 features; all classifiers must give a hypothesis in 18-dimensional space.

The proposed method is suitable for implementation. The recommendations of the British Diabetic Association (BDA) [35] are 80% sensitivity and 95% specificity. The test results of our method are close to those two standards.

5. Conclusion

In this paper, an active learning classifier is presented for further reducing diabetic retinopathy screening system cost. Classic researches did 5- or 10-fold cross-validation which implies that massive diagnosis results should be prepared beforehand. Unlike other state-of-the-art methods, we focus on further reducing cost. We use kernel extreme learning machine to deal with classification problem in high dimensional space. For solving overfitting problem brought by small training set, we adapt ensemble learning method. By using active learning with QBC, the ensemble-KELM learns from manual diagnosis result by necessary queries.

Our approach and other comparative classifiers had been validated on public diabetic retinopathy dataset. Kernel technique and bagging technique are also tested and analyzed. Empirical experiment shows that our approach can classify unlabeled retinal images with higher accuracies than other comparative classifiers, but the size of training dataset is much smaller than other comparative classifiers. With the consideration of implementation, the performance of our approach is close to the recommendations of the British Diabetic Association.

Competing Interests

The authors declare that they have no competing interests.

References

[1] A. N. Kollias and M. W. Ulbig, "Diabetic retinopathy: early diagnosis and effective treatment," *Deutsches Arzteblatt International*, vol. 107, no. 5, pp. 75–84, 2010.

[2] K. Venkatnarayan, J. Pboyle, and T. Jthompson, "Lifetime risk for diabetes mellitus in the United States," *Journal of the American Medical Association*, vol. 290, no. 14, pp. 1884–1890, 2003.

[3] M. Memon, S. Memon, and N. Bakhtnizamani, "Sight threatening diabetic retinopathy in type—2 diabetes mellitus," *Pakistan Journal of Ophthalmology*, vol. 30, no. 1, pp. 1–9, 2014.

[4] V. R. Driver, M. Fabbi, L. A. Lavery, and G. Gibbons, "The costs of diabetic foot: the economic case for the limb salvage team," *Journal of Vascular Surgery*, vol. 52, no. 3, pp. 17S–22S, 2010.

[5] R. Li, S. Sshrestha, and R. Dlipman, "Diabetes self-management education and training among privately insured persons with newly diagnosed diabetes—United States, 2011-2012," *Morbidity and Mortality Weekly Report*, vol. 63, no. 46, pp. 1045–1049, 2014.

[6] W. C. Chan, L. T. Lim, M. J. Quinn, F. A. Knox, D. McCance, and R. M. Best, "Management and outcome of sight-threatening diabetic retinopathy in pregnancy," *Eye*, vol. 18, no. 8, pp. 826–832, 2004.

[7] J. Cuadros and G. Bresnick, "EyePACS: an adaptable telemedicine system for diabetic retinopathy screening," *Journal of Diabetes Science and Technology*, vol. 3, no. 3, pp. 509–516, 2009.

[8] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool," *British Journal of Ophthalmology*, vol. 80, no. 11, pp. 940–944, 1996.

[9] P. Hscanlon, C. Pwilkinson, and S. Jaldington, *Screening for Diabetic Retinopathy*, 2009.

[10] C. Sinthanayothin, V. Kongbunkiat, and S. Phoojaruenchanachai, "Automated screening system for diabetic retinopathy," in *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis (ISPA '03)*, vol. 2, pp. 915–920, September 2003.

[11] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool," *The British Journal of Ophthalmology*, vol. 80, no. 11, pp. 940–944, 1996.

[12] G. Liew, C. A. Egan, A. Rudnicka et al., "Evaluation of automated software grading of diabetic retinopathy and comparison with manual image grading—an accuracy and cost effectiveness study," *Investigative Ophthalmology & Visual Science*, vol. 55, no. 13, p. 2293, 2014.

[13] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[14] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[15] K. G. M. M. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation," *Diabetic Medicine*, vol. 15, no. 7, pp. 539–553, 1998.

[16] S. Vijan, T. P. Hofer, and R. A. Hayward, "Cost-utility Analysis of screening intervals for diabetic retinopathy in patients with type 2 diabetes mellitus," *Journal of the American Medical Association*, vol. 283, no. 7, pp. 889–896, 2000.

[17] K. Shotliff and G. Duncan, "Diabetic retinopathy: summary of grading and management criteria," *Practical Diabetes International*, vol. 23, no. 9, pp. 418–420, 2006.

[18] http://archive.ics.uci.edu/ml/.

[19] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-Based Systems*, vol. 60, pp. 20–27, 2014.

[20] M. Dabramoff, J. Mreinhardt, and S. Rrussell, "Automated early detection of diabetic retinopathy," *Ophthalmology*, vol. 117, no. 6, pp. 1147–1154, 2010.

[21] A. D. Fleming, K. A. Goatman, S. Philip, J. A. Olson, and P. F. Sharp, "Automatic detection of retinal anatomy to assist diabetic retinopathy screening," *Physics in Medicine and Biology*, vol. 52, no. 2, pp. 331–345, 2007.

[22] A. Sopharak, B. Uyyanonvara, and S. Barman, "Automatic exudate detection for diabetic retinopathy screening," *ScienceAsia*, vol. 35, no. 1, pp. 80–88, 2009.

[23] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Neural Information Processing Systems*, 1995.

[24] H. Sseung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, Pittsburgh, Pa, USA, July 1992.

[25] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[26] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[27] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[28] G. Kovács and A. Hajdu, "Extraction of vascular system in retina images using averaged one-dependence estimators and orientation estimation in hidden markov random fields," in *Proceedings of the 8th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 693–696, IEEE, Chicago, Ill, USA, April 2011.

[29] B. Antal, A. Hajdu, Z. Maros-Szabó, Z. Török, A. Csutak, and T. Peto, "A two-phase decision support framework for the automatic screening of digital fundus images," *Journal of Computational Science*, vol. 3, no. 5, pp. 262–268, 2012.

[30] B. Antal, I. Lázár, and A. Hajdu, "An ensemble approach to improve microaneurysm candidate extraction," in *Signal Processing and Multimedia Applications*, vol. 222 of *Communications in Computer and Information Science*, pp. 378–394, Springer, Berlin, Germany, 2012.

[31] B. Nagy, B. Harangi, B. Antal, and A. Hajdu, "Ensemble-based exudate detection in color fundus images," in *Proceedings of the 7th International Symposium on Image and Signal Processing and Analysis (ISPA '11)*, pp. 700–703, Dubrovnik, Croatia, September 2011.

[32] B. Antal and A. Hajdu, "A stochastic approach to improve macula detection in retinal images," *Acta Cybernetica*, vol. 20, no. 1, pp. 5–15, 2011.

[33] R. J. Qureshi, L. Kovacs, B. Harangi, B. Nagy, T. Peto, and A. Hajdu, "Combining algorithms for automatic detection of optic disc and macula in fundus images," *Computer Vision and Image Understanding*, vol. 116, no. 1, pp. 138–145, 2012.

[34] C. Agurto, V. Murray, E. Barriga et al., "Multiscale AM-FM methods for diabetic retinopathy lesion detection," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 502–512, 2010.

[35] G. P. Leese, "Retinal photography screening for diabetic eye disease," Tech. Rep., British Diabetic Association, 1997.