Signal Processing ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Automatic cell nuclei segmentation and classification of breast cancer histopathology images

Pin Wang [a,*], Xianling Hu [a], Yongming Li [a,b], Qianqian Liu [a], Xinjian Zhu [c]

[a] College of Communication Engineering of Chongqing University, Chongqing 400030, PR China
[b] Department of Medical Imaging of College of Biomedical Engineering, Third Military Medical University, Chongqing 400038, PR China
[c] Research Institute of Field Surgery, Daping Hospital, Third Military Medical University, Chongqing 400038, PR China

ARTICLE INFO

Article history:
Received 28 June 2015
Received in revised form
4 November 2015
Accepted 15 November 2015

Keywords:
Breast cancer
Nuclei segmentation
Wavelet decomposition
Curvature scale space corner detection
Chain-like agent genetic algorithm
Support vector machine classification

ABSTRACT

Breast cancer is the leading type of malignant tumor observed in women and the effective treatment depends on its early diagnosis. Diagnosis from histopathological images remains the "gold standard" for breast cancer. The complexity of breast cell histopathology (BCH) images makes reliable segmentation and classification hard. In this paper, an automatic quantitative image analysis technique of BCH images is proposed. For the nuclei segmentation, top-bottom hat transform is applied to enhance image quality. Wavelet decomposition and multi-scale region-growing (WDMR) are combined to obtain regions of interest (ROIs) thereby realizing precise location. A double-strategy splitting model (DSSM) containing adaptive mathematical morphology and Curvature Scale Space (CSS) corner detection method is applied to split overlapped cells for better accuracy and robustness. For the classification of cell nuclei, 4 shape-based features and 138 textural features based on color spaces are extracted. Optimal feature set is obtained by support vector machine (SVM) with chain-like agent genetic algorithm (CAGA). The proposed method was tested on 68 BCH images containing more than 3600 cells. Experimental results show that the mean segmentation sensitivity was 91.53% ($\pm$4.05%) and specificity was 91.64% ($\pm$4.07%). The classification performance of normal and malignant cell images can achieve 96.19% ($\pm$0.31%) for accuracy, 99.05% ($\pm$0.27%) for sensitivity and 93.33% ($\pm$0.81%) for specificity.

1. Introduction

Breast cancer is the leading type of malignant tumor observed in women. Early detection of cancer is very important for successful treatment. Diagnosis from histopathological images remains the "gold standard" for breast cancer. Manually analyzing numerous biopsy slides by pathologist is labor intensive and has suboptimal reproducibility. Thanks to recent advances in digital pathology, the automatic image analysis method has the potential to overcome the subjective interpretation and reduce the workload. Computer-aided diagnosis (CADx) scheme is becoming an important tool to assist pathologist in breast cancer detection and diagnosis. The CADx scheme consists of two phases, which are segmentation phase and classification phase [1–3].

Segmentation of nuclei is an important first step towards automatic analysis of BCH images. Several algorithms for segmentation of nuclei in BCH images have been proposed. Most of them revolve around watershed segmentation, active contours, pixel classification or combination varieties, supplemented by different pre-processing and post-processing phases. Watershed segmentation is always improved by obtaining

* Corresponding author.
E-mail address: wangpin@cqu.edu.cn (P. Wang).

http://dx.doi.org/10.1016/j.sigpro.2015.11.011
0165-1684/

locations as markers that can mark the objects of interest and the background [4–8]. But these techniques suffer from over-segmentation and did not work well for the overlapping cells. The objective of active contours is to find a minimum energy fit of the moving contours, and the algorithm is always combined with a nuclei detection method. These kinds of methods initially define a large number of candidate regions and then select the ones that can present correctly segmented nuclei [8–13]. However, these models have some limitations in convergence, the optimization problem involved leads to uncertainty and poor stability of result of segmentation. Moreover, clustering based methods such as *K*-means [14] and unsupervised or supervised machine learning [12,15] have been applied for the segmentation of the cell nuclei of breast cancer. These methods require explicit prior knowledge of the image structure and the computational complexity is relatively high. Due to the high variability of the tissue appearance, reliable cell nuclei segmentation of BCH images is still a challenging task.

After nuclei are precisely segmented, classification phase is implemented. The most important aspects of the classification performance are the features extracted and the classification algorithms. Some researchers have studied the analysis of BCH images by proposing new features or considering different classification algorithms. Most of the extracted features are morphology-based and texture-based features. Some new features were extracted such as textural features using a critical exponent analysis (CEA) [16] and complex Daubechies wavelets [13,17], distribution-based features of nuclei [18], etc. Most classification algorithms utilized in literature revolve around Support Vector Machine (SVM) [19], *k*-Nearest Neighbor (*k*-NN) [13], Naive Bayes (NB) [18], fuzzy *c*-means (FCM) [20], neural network [19] or a combination of the above mentioned algorithm [18]. Some other classification approaches such as Decision Tree (DT) and partial least squares regression [21] have been applied for breast cell classification as well. Almost all the literature applied all extracted feature as the input of classifiers. However, feature selection can improve the classification accuracy and the reliability [18,22].

In this paper, an automatic CADx scheme of BCH images is proposed. For the nuclei segmentation of BCH images, top-bottom hat transform is applied to enhance grayscale image. Wavelet decomposition and multi-scale region-growing (WDMR) are combined to obtain regions of interest (ROIs), a double strategy splitting model (DSSM) containing adaptive mathematical morphology and Curvature Scale Space (CSS) corner detection method is applied to split overlapped cells for better accuracy and robustness. For the classification of the cell nuclei, 4 shape-based features and 138 textural

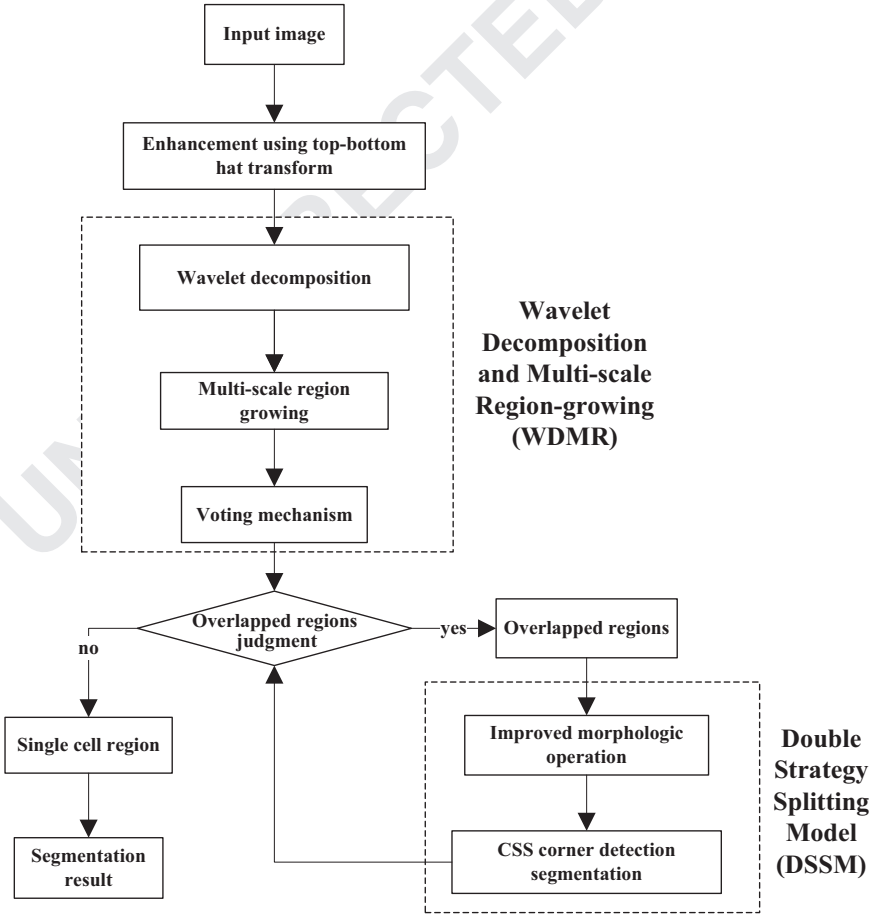

Fig.1. Block-diagram of the proposed segmentation method.

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

features based on color spaces are extracted as initial feature set. Then optimal feature set is selected by wrapper feature selection algorithm based on chain-like agent genetic algorithm (CAGA) [23] and SVM classifier to get high classification accuracy.

This paper is divided into five sections. Section 1 presents a review of the methods for cell nuclei segmentation and classification. Section 2 describes the acquisition progress of the images used for analyzing in this paper. Section 3 presents the methods in cell segmentation stage. Feature extraction, selection and classification algorithms are described in Section 4. Section 5 discusses the experiment results of segmentation and classification of BCH images. Finally the paper ends up with conclusions.

2. Tissue preparation and imaging

The first step of the tissue preparation process was formalin fixation and embedding in paraffin [24,25]. From the paraffin blocks, sections with a thickness of 4 μm were cut using a microtome (a high precision cutting instrument) and mounted on glass slides. Then the mounted sections were stained with H&E and cover-slipped in order to make the nuclei and cytoplasm visible. The histopathological images were acquired through digital camera adapted to an optical microscopy with 40 × magnification. Finally the captured RGB images are compressed with JPEG 2000.

3. Nuclei detection and segmentation

While all the subsequent processes are based on the segmentation results, it is very important that all the cells were isolated properly. An automatic segmentation method based on multi-scale region-growing with Double Strategy Splitting Model is proposed. The block-diagram with an overview of the proposed method is shown in Fig.1. This entire approach can be divided into three steps including pre-processing, ROIs extraction using WDMR, and overlapped cells isolation based on DSSM.

3.1. Pre-processing

Due to the digitized microscopic technique and compressing procedure, the background illumination of captured RGB images is usually not uniform. In order to gain better segmentation effect, top-bottom hat transform (TB_{hat}) is used to enhance the contrast between the cell nuclei and the background.

Two important steps of TB_{hat} transformation are top-hat (T_{hat}) and bottom-hat (B_{hat}) transformation. T_{hat} is usually used for relatively light objects in the dark background, while B_{hat} is on the contrary. Let f be a grayscale image, the enhancement of f can be defined by

$$\begin{cases} f^{enh} = TB_{hat} = f + T_{hat} - B_{hat} \\ T_{hat} = f - (f\ b) \\ B_{hat} = (f \bullet b) - f \end{cases} \quad (1)$$

where $f\ b$ and $f \bullet b$ performs morphological opening and closing operation on the grayscale image f with the structure element b.

3.2. ROIs extraction with WDMR algorithm

After pre-processing, the contrast between cells and background has been enhanced but they are not separated from each other. As the contrast and size of different cell nuclei are various, multi-resolution method is chosen to extract the ROIs. As it is difficult to see the detail information of low-contrast or small-size objects, we need to adopt high resolution observation. On the contrary, if the objects have high contrast or big size, low resolution observation is enough. If the contrasts or sizes of observed objects are various, it has great advantages to use different resolutions to observe different types of objects. The process is called multi-resolution observation. Wavelet transform is a typical and widely used category of multi-resolution image processing.

Wavelet transform is based on small waves called wavelets of varying frequency and limited duration. Discrete Wavelet Transform (DWT) can decompose a signal or image with expression of a set of translated and scaled version of a basis mother wavelet [26]. DWT employs a set of functions such as wavelet function $\psi(x)$ and scaling function $\varphi(x)$, which are functions of discrete variable $x = 0, 1, 2, \ldots, M-1$. As digitized image is often taken as two-dimensional discrete information, the DWT is extended to two-dimensional function. In two-dimension, a two-dimensional scaling function ($\varphi(x,y)$) and three two-dimensional wavelets ($\psi^H(x,y)$,$\psi^V(x,y)$,$\psi^D(x,y)$) are required. These wavelets can be named "directionally sensitive" wavelets for the measured functional variations (intensity or gray-level variations for images) along different directions. (The wavelet ψ^Hmeasures variations along columns, ψ^Vresponds to variations along rows, andψ^Dcorresponds to variations along diagonals). As Daubechies (dbs) has constructed a class of compactly supported real wavelets and has given a real solution for the wavelet equation [27], a special and simplest case of Daubechies wavelet named Haar wavelet which is also known as db1 is applied in this paper. The property of Haar wavelet that it is not continuous and differentiable can be an advantage for the analysis of signals or images with sudden transitions.

The enhanced grayscale image function $f^{enh}(x,y)$ of size $M \times N$ can be obtained. The corresponding discrete wavelet transform can be defined as

$$W_{\varphi}(j_0,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f^{enh}(x,y) \varphi_{j_0,m,n}(x,y) \quad (2)$$

$$W_{\psi}^{i}(j,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f^{enh}(x,y) \psi_{j,m,n}^{i}(x,y), i = \{H,V,D\} \quad (3)$$

where j_0 is an arbitrary starting scale and the $W_{\varphi}(j_0,m,n)$ coefficients define an approximation of $f^{enh}(x,y)$ at scale j_0. The $W_{\psi}^{i}(j,m,n)$ coefficients add horizontal, vertical and diagonal details for scales $j \geq j_0$. In order to detect the

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

edges of cells, unlike the traditional inverse discrete wavelet transform, directional details are reserved. The final result of wavelet transform is obtained by

$$f^{w}(x,y)=\frac{1}{\sqrt{MN}}\sum_{i=H,V,D}\sum_{j=j_0}^{\infty}\sum_{m}\sum_{n}W_{\psi}^{i}(j,m,n)\psi_{j,m,n}^{i}(x,y)$$

$$-\frac{1}{\sqrt{MN}}\sum_{m}\sum_{n}W_{\varphi}(j_0,m,n)\varphi_{j_0,m,n}(x,y) \quad (4)$$

The ROIs can be marked as many scattered points by $f^{w}(x,y)$ with few background locations. These points can be taken as the seed points of region growing algorithm which can group pixels or sub-regions into larger regions based on predefined criteria to extract integrated ROIs. The approach is started with a set of seed points. The pixels in the adjacent areas with the pixel of the seeds are compared. If they have similar features and meet certain criteria for growth, then put the pixel into the seed pixel area until all the pixels are scanned. As there still has some false seed points, morphological erosion by a radius R of disk structuring element (SE) is applied to get real cell marking points. The next step is to select criteria for region growing. In this paper we choose two criteria for a pixel to be put into the growth region. One is that the absolute gray-level difference between the adjacent pixel and the seed is less than threshold T. Another one is the pixel to be included in one of the regions is 8-connected. Both the criteria should be met at same time.

Determination of R and T is crucial for the performance of region growing. Due to the characteristics (various shape size, discontinuous edges, uneven internal gray levels, heavy adhesion, etc.) of breast histopathology images, a single pair of R and T cannot work well across different nuclei within one image. In this paper, three scales of R are adopted. A small one is used for marking all possible locations with no overlapping, a large one is for separating overlapping locations, and a median one is for complements. The T is determined by an iterative region growing method which begins with a small threshold T and stops when the gray and size similarity has a sharp enlargement. Single scale 1 can process uneven internal gray levels and discontinuous edges within the cell nuclei, but overlapped nuclei areas are aggregated together. Single scale 3 can separate the adhesion area appropriately but the problems of uneven internal gray levels and discontinuous edges are more severe. Scale2 is the compensation for the insufficient of scale 1 and scale 3. The

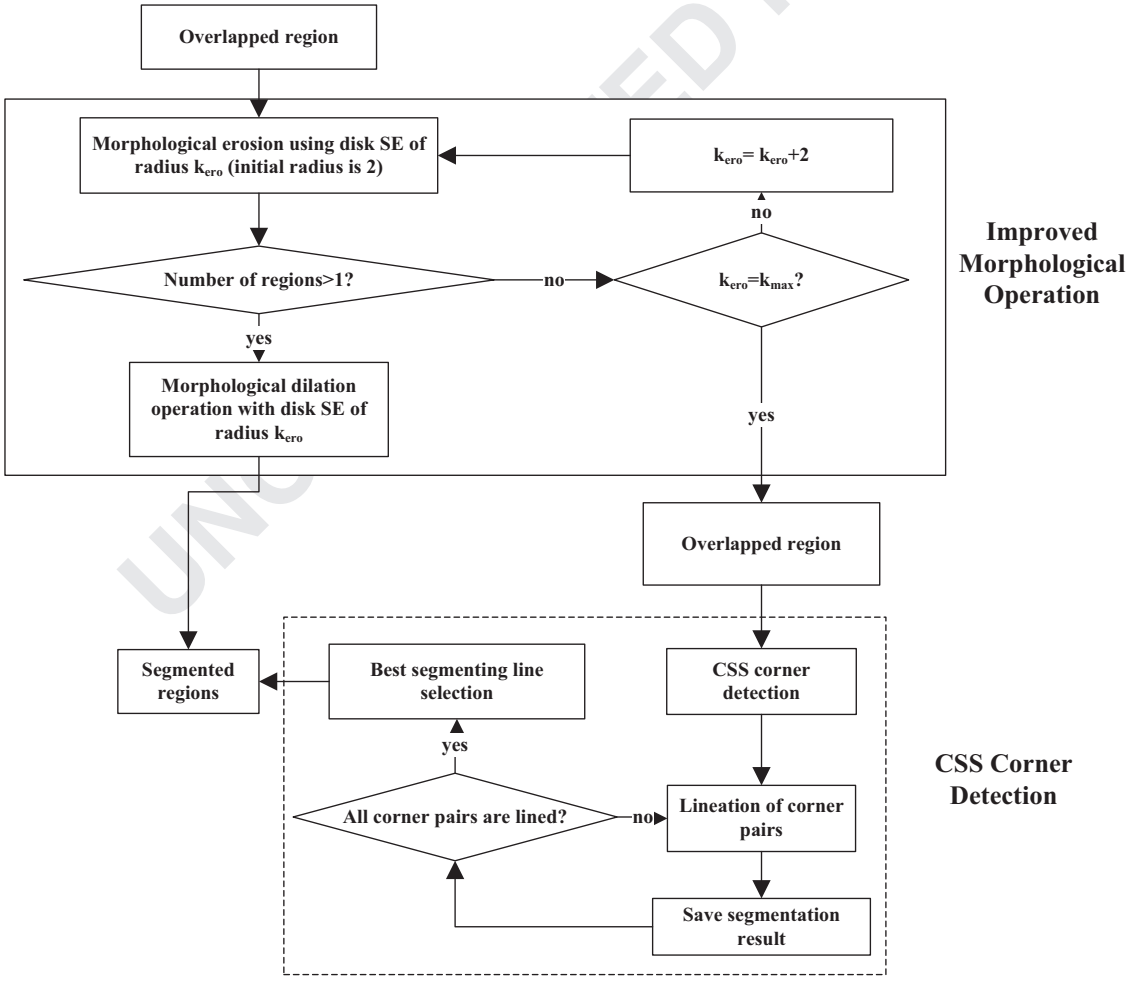

Fig.2. Block-diagram of DSSM.

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

algorithm adopted the three scales and can handle various and complex problems included in the images. Afterwards, the ROIs are chosen by a voting mechanism described as following:

Step 1: Calculate roundness circularity (R) and area (A) of every region of each scale.
Step 2: Save regions whose A is greater than 100 pixels for each scale as matrix f^{scale1}, f^{scale2} and f^{scale3} respectively.
Step 3: Choose one of regions in f^{scale1} and mark it by Reg_1. Mark the same location in f^{scale2} and f^{scale3} by Reg_2 and Reg_3. Choose the greatest R of the three regions as the best region and save it as matrix f^{ROIs}. Delete the three regions in the corresponding matrix.
Step 4: Repeat Step 2 and Step 3 until f^{scale1}, f^{scale2} and f^{scale3} are all empty.

3.3. Cell Isolation with DSSM model

The nuclear contour, in most cases, exhibited high radial symmetry. So the single and overlapped cell regions are separated by the following steps:

Step 1: Calculate roundness circularity (R) of every region in f^{ROIs}.
Step 2: Save regions whose R is greater than 0.7 as segmented out matrix f^{out} and others as unsegmented matrix f^{usg}.
Step 3: Compute the mean area of regions in f^{out} as the initial area threshold Ag.
Step 4: Compare the area of every region in f^{usg} with Ag, and save regions whose area bigger than Ag*1.5 as overlapped matrix f^{ove} and others as f^{out}.

According to the separated overlapped cell regions, double strategy splitting model (DSSM) combining adaptive mathematical morphology with Curvature Scale Space (CSS) corner detection method is applied to segment overlapped cells. The main process of this model is shown in Fig. 2.

As cells are convex and circular, there are concave corners along the overlapping borders of the cells. Morphological erosion operation can isolate overlapped cells by the concave corners and morphological dilation operation can reconstruct the eroded cells. Traditional morphological operation employs same size of SE on all cells which may lead to undesirable results for cells with different overlapping degree. An improved morphological operation is proposed for various overlapped cells. The proposed method chooses the appropriate size of SE adaptively to isolate the overlapped cells with less distortion. The concrete implementation process is shown as follows:

Step 1: Set the maximum and initial erosion radius (k_{max} and k_{ini}) of SE according to the radius of the biggest region in f^{out}.
Step 2: Employ erosion operation with disk SE of radius k_{ero} (initially set by k_{ini}) on overlapped region g^{ove} and obtain eroded matrix g^{ero}.
Step 3: If the number of connected regions is increased, go to step 4. Else, if $k_{ero}=k_{max}$, save g^{ove} as matrix f^{CSS} for further processing and go to end; else, let $k_{ero}=k_{ero}+2$, go to step 2.
Step 4: Reconstruct the regions of g^{ero} by using morphological dilation operation with disk SE of radius k_{ero} and save the reconstructed regions as f^{out}.

Some regions saved in f^{CSS} that cannot be isolated by the improved morphological operation need to be further segmented.

The cutting line of overlapped cells can be obtained from the concave corners detected. A number of corner detection methods have been proposed [28–30]. The curvature scale space (CSS) based corner detection technique is more suitable for recovering invariant geometric features of a planar curve [31–35]. This paper applies an improved multi-scale corner detector with dynamic region of support, which is based on CSS technique [33]. The detector first uses an adaptive local curvature threshold instead of a single global threshold and enhanced CSS methods. Second, the angles of corner candidates are checked in a dynamic region of support for eliminating falsely detected corners.

The enhanced CSS algorithm to detect corners of an image consists of six steps:

Step 1: Apply the Canny edge detector to the overlapped cells region and obtain binary edge contours.
Step 2: Fill the gaps in the contours and find the T-junctions.
Step 3: Compute curvature at a fixed low scale for each contour to retain all true corners.
Step 4: All of the curvature local maxima are considered as corner candidates. Then remove false corners to acquire correct corners.
Step 5: Track the corners from the highest scale to the lowest scale to improve localization.
Step 6: Compare the T-junction with other corners and remove one of them when they are very close.

For the step 4 among the corner candidates, although some points are detected as the local maximum, the values of their neighbor points are often very small. The local maxima points whose absolute curvatures are under its local threshold are eliminated. The adaptive threshold is given by

$$T(\mu,\sigma)=C\times\bar{\kappa}=1.5\times\frac{1}{L1+L2+1}\sum_{i=\mu-L2}^{\mu+L1}\kappa(i) \quad (5)$$

where C is a coefficient and $\bar{k}$ represents the mean curvature of the local region. If the curvature of the candidate corner is less than the threshold, it will be removed as false corner.

In general, a well-defined corner should have a relatively sharp angle. If we know the angle of each point on a curve, it would be easy to differentiate true corners from false corners. The corner checking criterion is given by

If $160^{\circ}\le\angle C_i\le 200^{\circ}$ then C_i is a False Corner
Else C_i is a True Corner.

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

$\angle C_i$ is given by

$$\angle C_i = \left| \tan^{-1}(\Delta Y1/\Delta X1) - \tan^{-1}(\Delta Y2/\Delta X2) \right| \quad (6)$$

where

$$\Delta X1 = \frac{1}{L1}\sum_{i=\mu+1}^{\mu+L1} X(i) - X(\mu) \Delta Y1 = \frac{1}{L1}\sum_{i=\mu+1}^{\mu+L1} Y(i) - Y(\mu) \quad (7)$$

$$\Delta X2 = \frac{1}{L2}\sum_{i=\mu-L2}^{\mu-1} X(i) - X(\mu) \Delta Y2 = \frac{1}{L2}\sum_{i=\mu-L2}^{\mu-1} Y(i) - Y(\mu) \quad (8)$$

Using this criteria, the false corner due to boundary noise and trivial details can be removed.

After corner candidates are extracted, cutting line is determined by several criteria:

1) The line should be as short as possible.
2) The corner pair to be lined should be in the opposite direction.
3) The segmented regions should be most similar to convex circle.

According to the criteria, the optimal cutting line is selected to reconstruct the contour of overlapped cells. The segmentation results based on CSS corner detection method are shown in Fig.3

As shown in second column in Fig. 3, the corners label the true concave points of overlapped cells and some false locations. The third column shows the optimal cutting lines selected by the criteria. Cell edges are delineated by pink contours on the original color images in fourth column for better observation. The segmentation results show that the optimal cutting lines marked the edges of cells precisely.

Fig.4 shows the segmentation results of cells with different overlapping degree using proposed DSSM algorithm. As shown in third column, the contours delineated by pink are well matched with true nuclei edges.

4. Classification

The efficient classification of cell nuclear requires the generation of meaningful features having very good discriminative ability. Morphometric, colorimetric, textural, and structural features have been used for feature selection and feature evaluation in previous studies to realize classification of breast cells. To obtain better classification result and time saving, shape-based features and textural features based on different color spaces are extracted and selected by CAGA. These discriminative features are proposed as input data for the SVM classifier.

4.1. Feature extraction

4.1.1. Shape features

There are certain differences in morphometric characteristics between normal and malignant cells. Four shape features are calculated including area, perimeter, eccentricity, and roundness circularity. Detailed information for these features is shown as follows:

Area: This is measured by calculating the number of pixels inside the cell region.

Perimeter: This is the total pixel number of the cell edge.

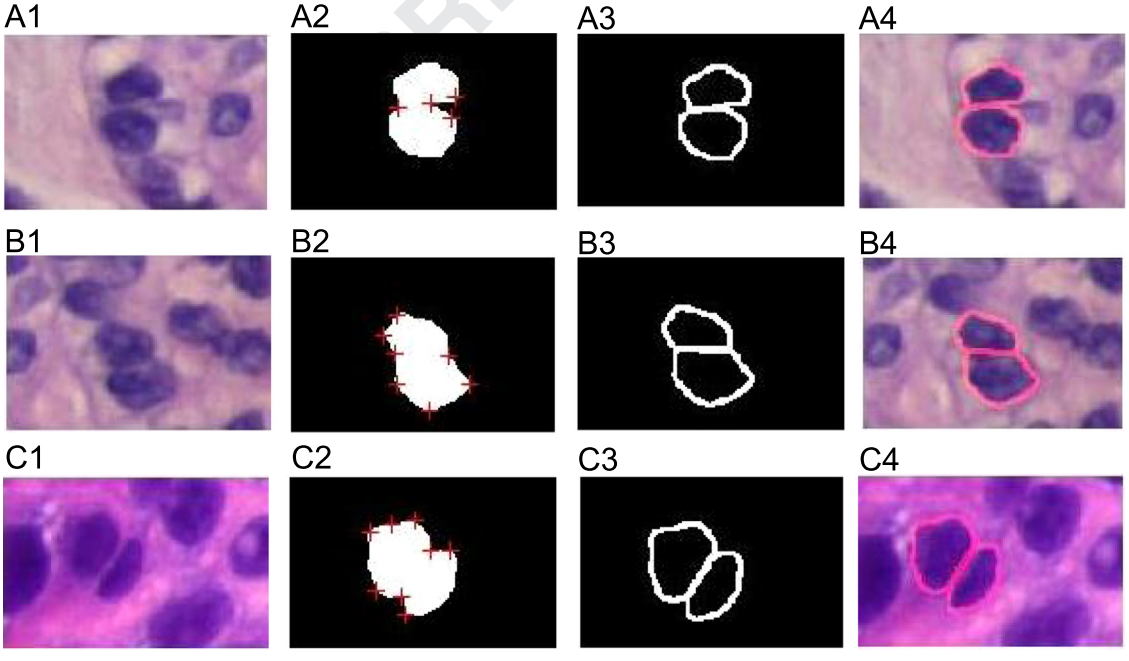

Fig.3. Separation results based on CSS corner detection method: (A1)–(C1) original color BCH images, (A2)–(C2) corresponding CSS corner detection results (red crossheads mark the corners), (A3)–(C3) optimal cutting lines, and (A4)–(C4) segmentation results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

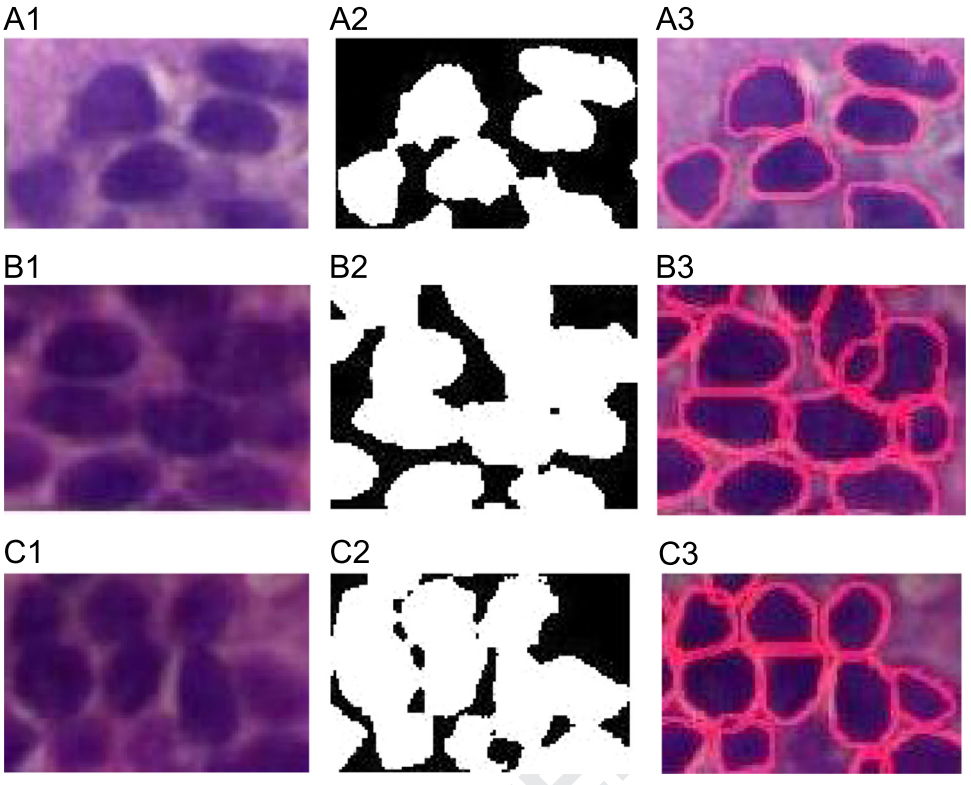

Fig.4. Segmentation results of overlapping cells using DSSM: (A1)–(C1) Original color BCH images, (A2)–(C2) extracted ROIs, and (A3)–(C3) segmentation results. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Eccentricity: This specifies the eccentricity of the ellipse that has the same second-moments as the cell region. The values of this feature vary between 0 and 1. A shape whose eccentricity is 0 is a circle while 1 is a line segment.

Roundness circularity: This is defined by $4\pi\times$ area/perimeter^2, which denotes the similarity between the cell region and a circle. The values vary between 0 and 1 and a circle's roundness circularity is equal to 1.

4.1.2. Textural features

In texture analysis, statistical methods can be used to analyze the spatial distribution of gray values, by computing local features at each point in the image, and deriving a set of statistics features from the distributions of the local features. It can describe the spatial arrangement of the pixel gray levels in a digitized microscopic nuclear image [36,37]. The selected first order statistical features for this study are mean and standard deviation of the gray level histogram. Let z be a random variable denoting gray levels and $p(z_i)$, $i=0,1,2,...,L-1$ be the corresponding histogram, where L is the number of distinct gray levels. The textural features obtained by statistical moments are: Mean value (the average gray level), Standard deviation (the 2nd moment about the mean value), Relative smoothness (measure of gray-level contrast), Third moment (skewness of the histogram), Uniformity(measure of the region gray consistency), and Entropy(measure of gray level variability).

Gray level components of seven color spaces, namely, RGB, HSV, HSI, YIQ, YCrCb, XYZ, Lab, combining with basic and enhanced gray-scaled components are employed to calculate their textural features.

4.2. Feature selection

The initial set of candidate feature consists of all the shape features and textural features. Almost all the literature used all extracted features to classify cells no matter whether it has significant difference between normal and malignant cells. Directly utilizing all features for classification may lead to more time consumption and poor performance [22]. A chain-like agent genetic algorithm (CAGA) proposed in [23] was applied to obtain the optimal subset of features for SVM classifier, and obtained satisfactory classification accuracy. The CAGA is implemented as follows:

Start

Ensemble n types of feature sets as to be selected feature sets $fea()$;

Procedure sample processing using $fea()$ so as to gain sample feature matrix$data_fea()_{m\times n}$. As for $data\_fea()_{m\times n}$, it is randomly divided into sample training feature matrix $data\_fea\_train()$ and sample testing matrix $data\_fea\_test()$;

A chain like agent structure sets up on account of CAGA.

According to no. i agent, $data\_fea\_test()$ and $data\_fea\_train()$ are being cut off separately, and then generate the corresponding sample training feature matrix $data\_fea\_train()_{ithAgent}$ and sample testing feature matrix $data\_fea\_test()_{ithAgent}$, then put $data\_fea\_train()_{ithAgent}$ into

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

classifier to operate training procedure, after this output the testing accuracy results as agent adaptations;

Finish all agent adaptation calculations;

Genetic operation (Neighborhood competition selection, self-adaption crossing, self-adaption mutation);

Verdict whether it meets determine condition; if so, go to the next step; if not, return;

Stop Output the optimal agent, which is the corresponding optimal feature subset with prior classification accuracy.

5. Results and discussion

The proposed automatic CADx scheme of BCH images is implemented as an algorithm by MATLAB. For the segmentation phase, in order to demonstrate the efficiency of the segmentation strategy, it is compared qualitatively and quantitatively with other methods. The segmentation method is applied on the 68 dataset images, and the performance is evaluated. For the classification phase, two different classification schemes are applied including image classification and patient classification. The classification results of the two schemes is discussed.

5.1. Nuclei Segmentation phase

The whole segmentation process is shown in Fig.5.

The parameters to be set for the experiment are listed as the following. Two-dimensional discrete Haar wavelet is chosen for wavelet decomposition. Three scales radius of multi-scale region-growing are 3, 5, and 7 pixels and the corresponding thresholds are automatically gotten by region growing algorithm criteria. For overlapped regions judgment, the basic roundness circularity is 0.7. The threshold of overlapped region is 1.5 times of Ag which is automatically calculated by segmented regions. During Adaptive mathematical morphology step, maximum erosion radius is determined by the radius of the biggest region in segmented regions, and the initial erosion radius is equal to 2. The coefficient of adaptive threshold is 1.5, and the false corner region is $160^{\circ} \le \angle C_i \le 200^{\circ}$ for Curvature Scale Space (CSS) corner detection.

The proposed segmentation method is employed on the captured BCH images by MATLAB. To evaluate the segmentation performance, the proposed method is compared with comparable methods from the published literature. The segmentation results are shown in Fig.6

As shown in row B and C, the marker-controlled watershed method based on adaptive H-minima algorithm [5] and FRST transform [4] markers selection can obtain good segmentation for cells which are significantly different from background. But both methods still suffer from over-segmentation and under-segmentation in complex images. Row D shows that *K*-means clustering combining with morphology operation [14] has better segmentation performance of homogeneous staining regions but suffers from many segmentation faults in uneven staining regions. The results of proposed method, as shown in row E, can reduce over-segmentation and under-segmentation efficiently and obtain precise contours of cells in complicated images.

Table 1 shows the quantitative comparison of segmentation performance of all 68 BCH images which are evaluated using visual inspection by experienced pathologists. There are 3663 cells in the images totally. As shown

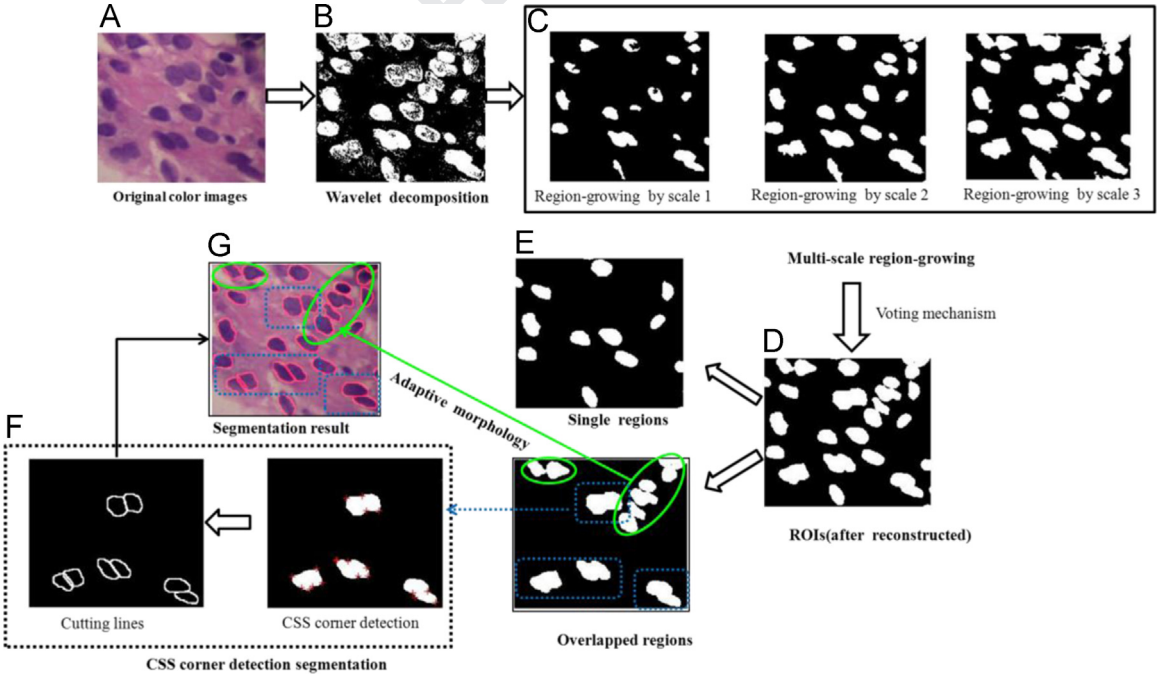

Fig.5. Segmentation process of one BCH image.

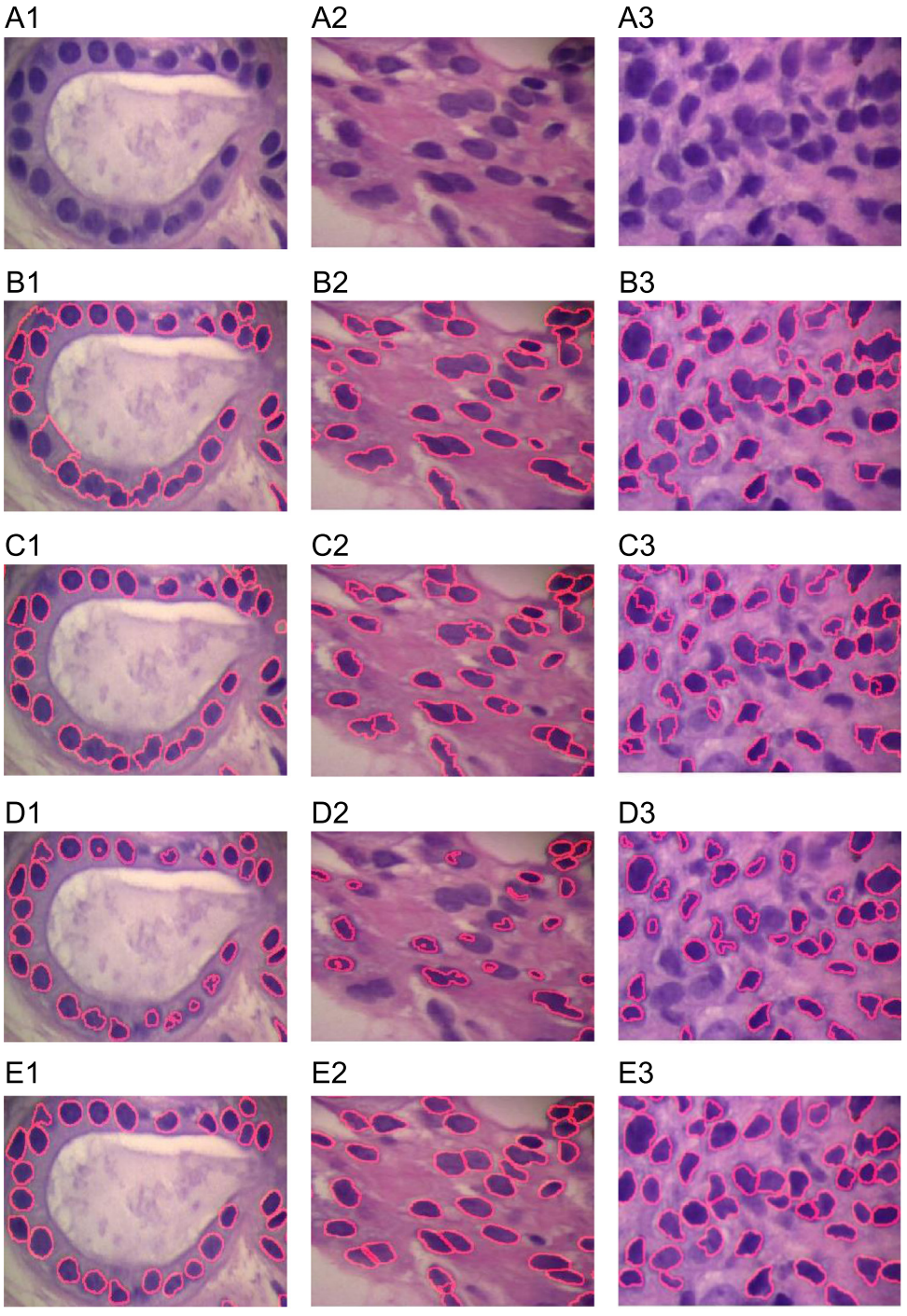

Fig.6. Segmentation results of breast histopathology images: (A1–A3) Original color breast cell images, (B1)–(B3) corresponding results of marker-controlled watershed based on adaptive H-minima algorithm for markers selection in [5], (C1)–(C3) corresponding results of marker-controlled watershed using FRST transform for markers selection in [4], (D1)–(D3) corresponding results of *K*-means clustering combining with morphology operation in [14], and (E1)–(E3) corresponding results of proposed WDMR&DSSM method.

in Table 1, it is clear that the segmentation performance of proposed method is better than those of other methods in terms of sensitivity and specificity. The mean values are above 90%, and meet the clinical requirement. The stand deviation is about 4%, and it means the performance is very stable. For the other three methods, due to the high

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

Table 1
Quantitative comparison results of segmentation performance.

Method	True segmented cells	Wrong segmented cells	Total cells	Sensitivity	Specificity
Ref. [5]	1992	1750	3663	55.27% ± 9.65%	52.91% ± 10.11%
Ref. [4]	2209	1698	3663	60.18% ± 11.43%	53.20% ± 11.98%
Ref. [14]	1755	1953	3663	47.69% ± 14.60%	45.99% ± 16.31%
Proposed	3349	306	3663	91.53% ± 4.05%	91.64% ± 4.07%

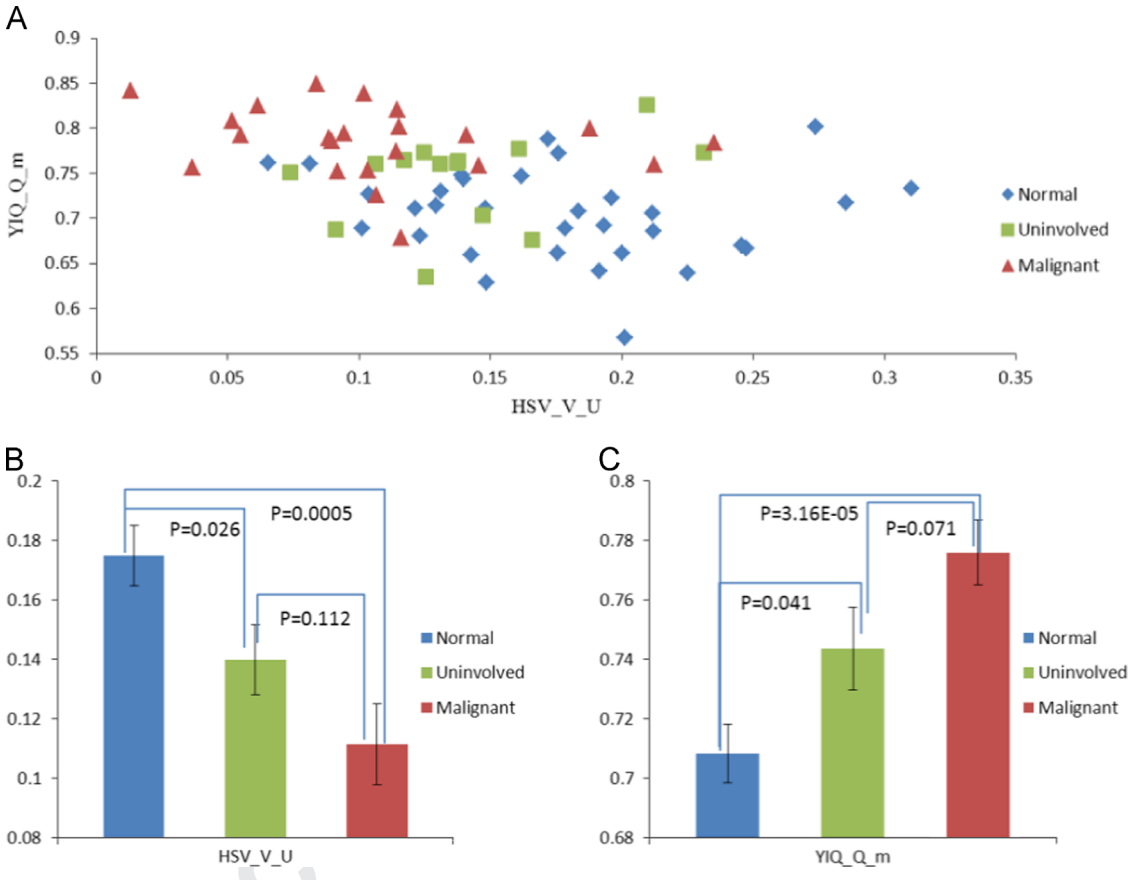

Fig.7. (a) Scatter plot of the average HSV_V_U and YIQ_Q_m of normal cell images (blue diamond), uninvolved cell images (green square), and malignant cell images (red triangle). Each point corresponds to a single image. (b) The bar graph of statistical analysis of HSV_V_U and YIQ_Q_m for three types of images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

variability of the tissue appearance the standard deviation is high and the segmentation performance is not satisfying.

5.2. Classification phase

The selection of classifiers for medical diagnosis applications depends on the sample size and the selection of relevant features from the data set. In this paper, the images are categorized with three main types, namely, (a) normal (normal cells from healthy patients), (b) uninvolved (i.e., cells from patients with invasive cancer that were called normal by pathological diagnosis), and (c) malignant (i.e. cells that were histologically called malignant from patients with invasive breast cancer) cells. In our study, we considered the use of two different classification schemes (Schemes 1 and 2). Q3

Scheme 1 Image classification scheme: In this approach, the mean values of the features for each image are calculated. Then classify images as normal, uninvolved or malignant images.

Scheme 2 Patient classification scheme: In this scheme, mean values of the features for slide of each patient are calculate, which means the information from all the images belonging to a given patient is collected.

For normal vs. malignant and normal vs. uninvolved, two features are selected (two-sided p-value: <0.05). They are HSV_V_U and YIQ_Q_m which present the uniformity of V component of HSV color space and the Q

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

Table 2
Classification performance of image scheme.

	Accuracy	Sensitivity	Specificity
Normal vs. malignant	96.19% ± 0.31%	99.05% ± 0.27%	93.33% ± 0.81%
Normal vs. uninvolved	93.81% ± 0.80%	95.24% ± 1.88%	92.38% ± 1.09%
Normal vs. malignant + uninvolved	95.83% ± 0.10%	98.81% ± 0.07%	92.86% ± 0.42%

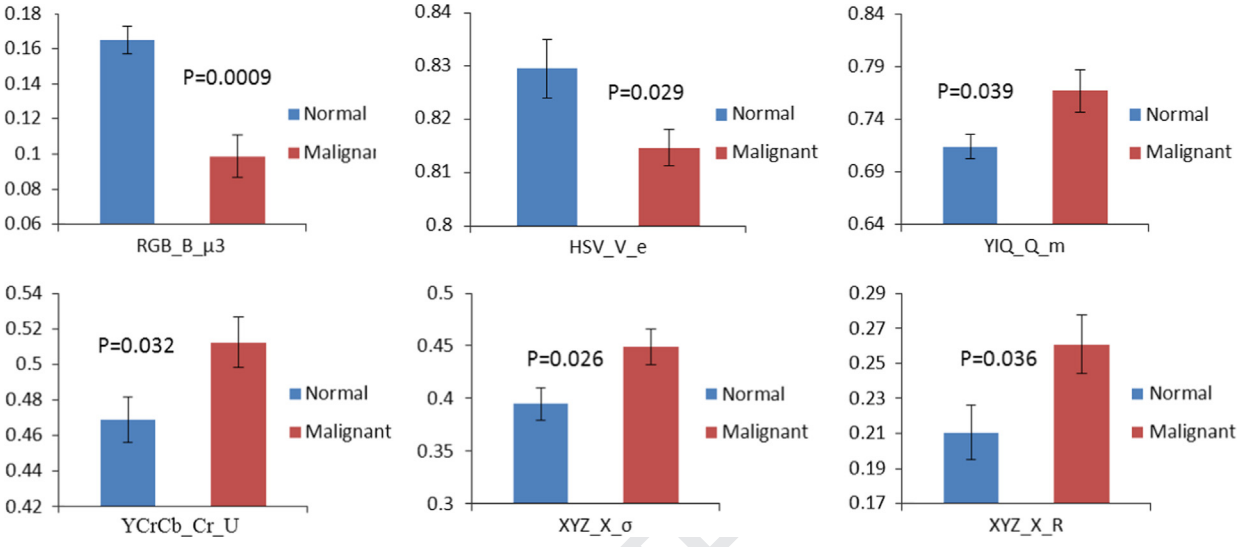

Fig.8. Bar graphs of features with statistical significance for normal and cancer cases.

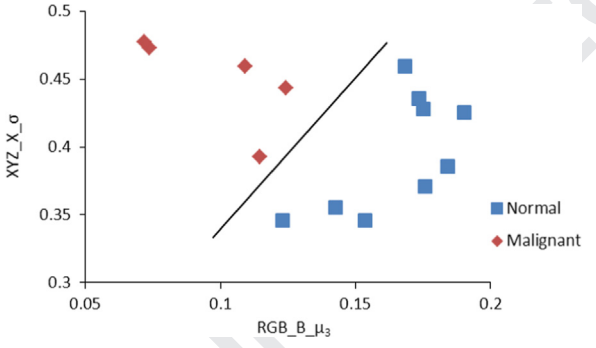

Fig.9. Scatter plot of the individual-patients based characteristic features.

component of YIQ color space. The scatter plot and bar graph of these three types of images are shown in Fig.7.

Fig. 7(a) shows a scatter plot of the average HSV_V_U and YIQ_Q_m of all images in each group: (1) normal cell images; (2) uninvolved cell images; and (3) malignant cell images. Despite the intrinsic variations among different nuclei, the malignant images and normal images are well separated. Importantly, the majority of uninvolved images in a cancer patient are clearly distinguishable from normal images from healthy patient.

Fig. 7(b) shows the statistical analysis of two features from all cells in 68 images. The HSV_V_U was significantly decreased (p-value < 0.001) in malignant cancer cells compared with that in normal cells, while the YIQ_Q_m was significantly increase. The same features can also distinguish uninvolved cells in cancer patients from cells in healthy patients significantly (p-value < 0.05). On the other hand, the two features does not present any statistical difference between uninvolved cells and malignant cells (p-value $= 0.1$ and 0.07). It suggests that although uninvolved cells are diagnosed as normal, they could alter cellular processes during neoplastic transformation such as cell growth, repair, apoptosis, and cell cycle regulation and result in the observed changes in our texture markers.

To obtain precise classification results, the classification algorithm based on CAGA combining with SVM classifier is applied. This method firstly employs SVM classifier on all features with kernel function polynomial, thus obtains initial feature set and accuracy. Then CAGA is applied to select optimal feature set with best accuracy. Finally the best accuracy and corresponding optimal feature set are stored. To get convincing classification results, average data of 30 times classification is required.

Classification results of image scheme are shown in Table 2. There are 32 normal cell images, 14 uninvolved cell images and 22 malignant cell images. The implement of classification is categorized into 3 groups, namely, normal vs. malignant images, normal vs. uninvolved images, and normal vs. malignant add with uninvolved images. Every group is repeated for 30 times with 30 iterations. As shown in Table 2, the proposed classification algorithm can effectively classify normal, uninvolved and malignant cell images. The accuracy of normal vs. malignant can reach an average of 96.19% and there is also a significant

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

classification performance of normal vs. uninvolved cell images.

Next step, the features of all images from each patient is calculated for the patient classification scheme. For the data set, there are 9 normal patients and 5 cancerous patients. There are six features significantly (p-value < 0.05) including RGB_B_μ3 (third moment of B component of RGB), HSV_V_e (entropy of V component of HSV), YIQ_Q_m (mean value of Q component of YIQ), YCrCb_Cr_U (uniformity of Cr component of YCrCb), XYZ_X_σ (standard deviation of X component of XYZ) and XYZ_X_R (relative smoothness of X component of XYZ). The bar graphs of the features with statistical significance for the cases are shown in Fig.8.

As shown in Fig.8, RGB_B_μ_3 and XYZ_X_σ have a higher statistical significance (p- value < 0.03). Scatter plot based on the feature RGB_B_μ_3 and XYZ_X_σis shown in Fig. 9. It is evident that there is a clear linear separability between the normal patients and cancer patients group, as indicated by the solid line in Fig. 9.

6. Conclusions

This paper describes a computer-aided diagnosis system for quantitative analyzing the BCH images. A new fully automatic segmentation method is developed for detection and segmentation of cell nuclei. The proposed method utilizes wavelet transform and multi-scale region growing to locate ROIs. Then an adaptive morphological operation combining with CSS corner detection algorithm is applied for separating overlapping cells. Compared with other cell image segmentation method, the segmentation results of proposed method have higher segmentation accuracy. Shape and texture features of cell nuclear are extracted for classification. Feature selection algorithm based on CAGA and SVM is applied to get the optimal feature set. The classification results show that the selected features can significantly distinguish normal and malignant images and has high classification accuracy. We demonstrate the promise of this technique for improving breast cancer diagnosis through characterizing the features of cell nuclei, especially its potential in detecting cancer from histopathological normal-appearing cells. In the future, the proposed method needs to be tested in a larger data set in order to further evaluate the robustness of the method.

Acknowledgment

Q4 This research is funded by National Natural Science Foundation of China NSFC (Nos. 61108086, 61171089, and Q5 11304382), the Natural Science Foundation of Chongqing (cstc2012jjA40015), Chongqing City Science and Technology Plan (cstc2012gg-yyjs0572), Fundamental Research Funds for the Central Universities (CDJZR12160011, CDJZR13160008, and CDJZR155507), The China Postdoctoral Science Foundation (2013M532153) and the Chongqing Postdoctoral Science Special Foundation of China.

References

[1] F. Bray, J. Ren, E. Masuyer, J. Ferlay, Estimates of global cancer prevalence for 27 sites in the adult population in 2008, Int. J. Cancer 132 (5) (2012) 1133–1145.
[2] M. May, A better lens on disease: computerized pathology slides may help doctors make faster and more accurate diagnoses, Sci. Am. 302 (2010) 74–77.
[3] N. Stathonikos, M. Veta, A. Huisman, et al., Going fully digital: perspective of a Dutch academic pathology lab, J. Pathol. Inform. (2013) 4.
[4] M. Veta, P.J. van Diest, R. Kornegoor, A. Huisman, M.A. Viergever, et al., Automatic nuclei segmentation in H&E stained breast cancer histopathology images, PLoS One 8 (7) (2013) e70221.
[5] J. Cheng, J.C. Rajapakse, Segmentation of clustered nuclei with shape markers and marking function, IEEE Trans. Biomed. Eng. 56 (3) (2009) 741–748.
[6] X. Yang, H. Li, X. Zhou, Nuclei segmentation using marker controlled watershed, tracking using mean-shift, and kalman filter intimelapse microscopy, IEEE Trans. Circuits Syst. I: Regul. Pap. 53 (2006) 2405–2414.
[7] M.G. Yasmeen, M.B. Bassant, H.Z. Hala, I.R. Mohamed, Automated cell nuclei segmentation for breast fine needle aspiration cytology, Signal Process. 93 (10) (2013) 2804–2816.
[8] A. Mouelhi, M. Sayadi, F. Fnaiech, K. Mrad, K.B. Romdhane, Automatic image segmentation of nuclear stained breast tissue sections using color active contour model and an improved watershed method, Biomed. Signal Process. Control 8 (5) (2013) 421–436.
[9] S. Ali, A. Madabhushi, An Integrated region-, boundary-, shape based active contour for multiple object overlap resolution in histological imagery, IEEE Trans. Med. Imaging 31 (7) (2012) 1448–1460.
[10] H. Fatakdawala, J. Xu, A. Basavanhally, G. Bhanot, S. Ganesan, et al., Expectation-maximization-driven geodesic active contour with overlap resolution (EMaGACOR): application to lymphocyte segmentation on breast cancer histopathology, IEEE Trans. Biomed. Eng. 57 (7) (2010) 1676–1689.
[11] X. Qi, F. Xing, D.J. Foran, L. Yang, Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set, IEEE Trans. Biomed. Eng. 59 (3) (2012) 754–765.
[12] J.P. Vink, M.B. Van Leeuwen, C.H.M. Van Deurzen, H.G. De, Efficient nucleus detector in histopathology images, J. Microsc. 249 (2) (2013) 124–135.
[13] P. Bamford, B. Lovell, Unsupervised cell nucleus segmentation with active contours, Signal Process. 71 (2) (1998) 203–213.
[14] N.S. Issac, P. Palanisamy, K. Sujathan, E. Bengtsson, Analysis of nuclei textures of fine needle aspirated cytology images for breast cancer diagnosis using Complex Daubechies wavelets, Signal Process. 93 (10) (2013) 2828–2837.
[15] C. Jung, C. Kim, S.W. Chae, S. Oh, Unsupervised segmentation of overlapped nuclei using bayesian classification, IEEE Trans. Biomed. Eng. 57 (12) (2010) 2825–2832.
[16] A. Phinyomark, S. Jitaree, P. Phukpattaranont, P. Boonyapiphat, Texture analysis of breast cancer cells in microscopic images using critical exponent analysis method, Procedia Eng. 32 (2012) 232–238.
[17] S.I. Niwas, P. Palanisamy, K. Sujathan, Wavelet based feature extraction method for breast cancer cytology images, IEEE Symp. Ind. Electron. Appl. (2010) 686–690.
[18] P. Filipczuk, T. Fevens, A. Krzyzak, R. Monczak, Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies, IEEE Trans. Med. Imaging 32 (12) (2013) 2169–2178.
[19] Y.M. George, H.H. Zayed, M.I. Roushdy, B.M. Elbagoury, Remote computer-aided breast cancer detection and diagnosis system based on cytological images, IEEE Syst. J. 8 (3) (2014) 949–964.
[20] J. Malek, A. Sebri, S. Mabrouk, et al., Automated breast cancer diagnosis based on GVF-snake segmentation, wavelet features extraction and fuzzy classification, J. Signal Process. Syst. 55 (1–3) (2009) 49–66.
[21] X. Xiong, Y. Kim, Y. Baek, et al., Analysis of breast cancer using data mining & statistical techniques, Softw. Eng. Artif. Intell. Netw. Parallel/Distrib. Comput. A (2005) 82–87.
[22] M. Kudo, J. Sklansky, Comparison of algorithms that select features for pattern classifiers, Pattern Recognit. 33 (1) (2000) 25–41.
[23] Y.M. Li, X.P. Zeng, L. Han, P. Wang, Two coding based adaptive parallel co-genetic algorithm with double agents structure, Eng. Appl. Artif. Intell. 23 (4) (2010) 526–542.
[24] A.C. Ruifrok, D.A. Johnston, Quantification of histochemical staining by color deconvolution, Anal. Quant. Cytol. Histol. 23 (4) (2001) 291–299.

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011

[25] M. Veta, P.W.P. Josien, J.D. Paul, M.A. Viergever, Breast cancer histopathology image analysis: a review, IEEE Trans. Biomed. Eng. 61 (5) (2014) 1400–1411.
[26] J.C. Goswami, A.K. Chan, Fundamentals of Wavelets: Theory, Algorithms, and Applications, xviii,359, Wiley & Sons, 2011.
[27] A. Khare, U.S. Tiwary, Daubechies complex wavelet transform based technique for denoising of medical images, Int. J. Image Graph. 7 (4) (2007) 663–687.
[28] Y. Liu, M. Hou, X. Rao, et al., A steady corner detection of gray level images based on improved Harris algorithm, in: IEEE International Conference on Networking, Sensing andControl, 2008, pp. 708–713.
[29] C.H. Yeh, Wavelet-based corner detection using eigenvectors of covariance matrices, Pattern Recognit. Lett. 24 (3) (2003) 2797–2806 .
[30] X. Gao, F. Sattar, R. Venkateswarlu, Multiscale corner detection of gray level images based on Log-Gabor wavelet transform, IEEE Trans. Circuits Syst. Video Technol. 17 (7) (2007) 868–875.
[31] F. Mokhtarian, R. Suomela, Robust image corner detection through curvature scale space, IEEE Trans. Pattern Anal. Mach. Intell. 20 (12) (1998) 1376–1381.
[32] Z. Li ,Y. Shen, A robust corner detector based on curvature scale space and Harris, in: IEEE International Conference on Image Analysis and Signal Processing, 2011, pp. 223–226.
[33] X.C. He, N.H.C. Yung, Curvature scale space corner detector with adaptive threshold and dynamic region of support, in: IEEE International Conference on Pattern Recognition, Computer Society, 2004, pp. 791–794.
[34] B. Zhong, W. Liao, Direct curvature scale space: theory and corner detection, IEEE Trans. Pattern Anal. Mach. Intell. 29 (3) (2007) 508–512.
[35] M. Awrangjeb, G. Lu, M. Murshed, An affine resilient curvature scale-space corner detector, in: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, 2007, pp. I-1233–I-1236.
[36] K. Rodenacker, E. Bengtsson, A feature set for cytometry on digitized microscopic images, Anal. Cell. Pathol. 25 (1) (2003) 1–36.
[37] B. Nielsen, Statistical nuclear texture analysis in cancer research: a review of methods and applications, Crit. Rev. Oncog. 14 (2) (2008) 89–164.

Please cite this article as: P. Wang, et al., Automatic cell nuclei segmentation and classification of breast cancer histopathology images, Signal Processing (2015), http://dx.doi.org/10.1016/j.sigpro.2015.11.011