NUCLEI SEGMENTATION FOR COMPUTER–AIDED DIAGNOSIS OF BREAST CANCER

MAREK KOWAL, PAWEŁ FILIPCZUK

Institute of Control and Computation Engineering
University of Zielona Góra, ul. Podgórska 50, 65-246 Zielona Góra, Poland
e-mail: \{M.Kowal,P.Filipczyk\}@issi.uz.zgora.pl

Breast cancer is the most common cancer among women. The effectiveness of treatment depends on early detection of the disease. Computer-aided diagnosis plays an increasingly important role in this field. Particularly, digital pathology has recently become of interest to a growing number of scientists. This work reports on advances in computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. The task at hand is to classify those as either benign or malignant. We propose a robust segmentation procedure giving satisfactory nuclei separation even when they are densely clustered in the image. Firstly, we determine centers of the nuclei using conditional erosion. The erosion is performed on a binary mask obtained with the use of adaptive thresholding in grayscale and clustering in a color space. Then, we use the multi-label fast marching algorithm initialized with the centers to obtain the final segmentation. A set of 84 features extracted from the nuclei is used in the classification by three different classifiers. The approach was tested on 450 microscopic images of fine needle biopsies obtained from patients of the Regional Hospital in Zielona Góra, Poland. The classification accuracy presented in this paper reaches 100%, which shows that a medical decision support system based on our method would provide accurate diagnostic information.

Keywords: computer-aided diagnosis, breast cancer, pattern analysis, fast marching.

1. Introduction

According to the International Agency for Research on Cancer and the National Cancer Registry in Poland, breast cancer is the most common cancer among women. In 2008, there were 1,384,155 diagnosed cases of breast cancer and 458,503 deaths caused by the disease worldwide (Ferlay et al., 2010; Bray et al., 2012). In 2010, there were 15,784 diagnosed cases among Polish women, and 5,226 resulted in death (National Cancer Registry in Poland, 2012). There has also been an increase in the incidence of breast cancer by 3–4% a year since the 1980s. The effectiveness of treatment largely depends on timely detection of the disease.

An important and often used diagnostic method is the so-called triple-test, which is based on three medical examinations and is used to achieve high confidence in the diagnosis. The triple-test includes self examination (palpation), mammography or ultrasonography imaging, and Fine Needle Biopsy (FNB) (Underwood, 1987). FNB is an examination that consists in obtaining material directly from the tumor. The collected material is then examined under a microscope to determine the prevalence of cancer cells. This approach requires extensive knowledge and experience of the cytologist responsible for the diagnosis. Automatic morphometric diagnosis can help make the results objective and assist inexperienced specialists. It also allows screening on a large scale where only difficult and uncertain cases would require further examination by the specialist. Along with the development of advanced vision systems and computer science, quantitative cytopathology has become a useful method for detection of diseases, infections as well as many other disorders (Gurcan et al., 2009; Śmietański et al., 2010; Hassan et al., 2010).

Recently a large amount of studies has been conducted on computer-aided breast cancer diagnosis based on mammography, ultrasonography and microwave imaging (Moon et al., 2011; Mohanty et al., 2013; Cheng et al., 2010; Moghbel and Mashohor, 2013; Verma et al., 2010; Li et al., 2013; Xu et al., 2012; Ganesan et al., 2013; Grzegorczyk et al., 2012; Kirshin et al., 2013; Nikolova, 2011). Some of the proposed solutions have been successfully used in hospitals and research centers (Birdwell et al., 2005; Butler et al., 2004; Cupples et al., 2004; Dean and Ilvento, 2006; Destounis et al., 2004; Doi, 2005; Gi-
Computer-aided diagnosis based on cytological images has not been explored so deeply yet. However, in recent years, there has been observed an increased interest in this area (Fabregue et al., 2011; Fatakdawala et al., 2010; Basavanahally et al., 2013; Gianatti et al., 2010; Bandyopadhyay et al., 2010; Lopez et al., 2009; Christel et al., 2011; Fuchs and Buhmann, 2011).

Many researchers have studied the segmentation of cytological images of breast tumors, proposed new features or tested the classification algorithms (Muniandy and Stanslas, 2008; Yasmeen et al., 2013; Mat-Isa et al., 2007; Cruz-Ramirez et al., 2009; Uneyli, 2007; Polat and Gunes, 2007; Jeleni et al., 2010; Niwas et al., 2013; Malek et al., 2009; Xiong et al., 2005). However, a few of these researchers have tested the efficiency of their methodology in a comprehensive computerized breast cancer classification system. Jeleni et al. (2010) presented an approach based on the level set segmentation method. Classification efficiency was tested on 110 (44 malignant, 66 benign) images with results reaching 82.6%. Niwas et al. (2013) presented a method based on the analysis of nuclei texture using a wavelet transform. Classification efficiency with the k-nearest neighbor algorithm on 645 (311 malignant, 334 benign) images reached 93.9%. Another approach was presented by Malek et al. (2009). They used active contours to segment nuclei and classified 200 (80 malignant, 120 benign) images using the fuzzy c-means algorithm, achieving 95% efficiency. Breast cancer diagnosis was also discussed by Xiong et al. (2005). Partial least squares regression was used to classify 699 (241 malignant, 458 benign) images, yielding 96.57% efficiency. However, the authors did not describe the segmentation method used to extract nuclei.

This paper presents recent progress in the development of a comprehensive fully automatic breast cancer diagnostic system based on analysis of cytological images of FNB material. The task at hand is to classify a case as benign or malignant. This is done by using morphometric, textural and topological features of nuclei isolated from microscopic images of the tumor.

In previous work we used a segmentation method based on the combination of adaptive thresholding in grayscale and clustering in the color space (Filipczuk et al., 2011a; 2011b; Kowal et al., 2011). Although this approach gives satisfactory results, it leaves room for improvement. The main disadvantage of the previous method is that it can generate a large number of objects containing two or more nuclei merged together. This happens when nuclei form dense three-dimensional clusters and overlap each other. To overcome this problem, we propose a new robust segmentation procedure. The binary image obtained using adaptive thresholding and clustering is conditionally eroded. As a result, centers of the nuclei are determined and used to initialize the multi-label fast marching algorithm. While relatively rarely mentioned in the literature, the algorithm gives highly satisfactory results for segmentation of cytological images.

From the selected nuclei, we extract a set of 84 features which are then tested by three different classifiers. The system scheme is presented in Fig. 1. The entire approach was tested on real medical images obtained from patients of the Regional Hospital in Zielona Góra, Poland. As shown later in this paper, the classification accuracy reached 100%. The results demonstrate that a computerized medical diagnosis system based on our method would be effective and can provide valuable, accurate diagnostic information.

The paper is divided into five sections. Section 1 presents an introduction into breast cancer diagnosis. Section 2 describes the acquisition process of the medical images used for testing. Segmentation, feature extraction and classification are described in Section 3. Section 4 shows the experimental results obtained by the proposed method. The paper ends with conclusions.

2. Material

All methods presented in this work were tested on real medical data. For this purpose, 450 images were collected from 50 patients (25 benign and 25 malignant). Each patient is represented by 9 images. The number of images was recommended by the specialists from the hospital and allows correct diagnosis by a pathologist.

The cytological material was obtained by FNB from patients of the Regional Hospital in Zielona Góra, Poland. Biopsies without aspiration were performed under the control of an ultrasonograph with a 0.5 mm diameter needle. Smears from the material were fixed in spray fixative (Cellfix by Shandon) and dyed with hematoxylin and eosin (H&E). The time between preparation of smears and their preservation in fixative never exceeded 3 seconds. All cancers were histologically confirmed and all patients with the benign disease were either biopsied or followed for a year.

The images were recorded by a Sony CCD Iris color video camera mounted atop an Axiophot microscope. The slides were projected into the camera with 160× objective and 2.5× ocular, giving together an enlargement of 400×. Images are BMP files, 704×576 pixels, 8 bit/channel RGB.
A properly prepared slide contains cells from the tumor and sometimes other cells, i.e., red blood cells or fat cells. Differentiation of malignant and benign cells is based on several important characteristics associated with cell nuclei and cytoplasm. Pathologists usually pay great attention to the size, shape and distribution of the nuclei, distribution of the chromatin in the nucleus and ratio of the area of cytoplasm to the area of the nucleus. Cancerous cells usually have a larger and irregular nuclei, in contrast to the smaller and uniform nuclei in benign cells. Furthermore, cancerous cells tend to break up during smear preparation and therefore have a tendency to form a three dimensional structure called nests or to spread out on the whole slide. Benign cells are usually strong enough not to tear while a physician is performing smear and they usually form single-layered structures. Chromatin in cancerous cells frequently forms lumps, while in benign cells chromatin is usually homogenous. Unfortunately, classification of malignant and benign cells is a very difficult task and requires a lot of experience because all these differences do not always occur at the same time and are usually very subtle. Therefore, the pathologist usually needs to investigate several areas of the slide in order to deliver a diagnosis.

Automated classification of tumor malignancy requires isolating relevant nuclei from the rest of the image. In the literature, many different approaches already have been proposed to extract cells or nuclei from microscope images (Al-Kofahi et al., 2010; Clocksin, 2003; Cloppet and Boucher, 2008; Krawczyk et al., 2012; Jeleń et al., 2010; Kowal and Korbič, 2010; Marciniak et al., 2005; Obuchowicz et al., 2008; Gochlawski et al., 2012). Unfortunately, reliable cell or nuclei segmentation is a challenging task. The appearance and the quality of the slide strongly depends on the experience of the person who prepared it. Distribution of the cells on the slide is random and can be completely different on the consecutive slides. Cells often cluster and overlap together, their boundaries are not clear and their interiors are not uniform (Fig. 2(a)). Moreover, attempts to generalize segmentation approaches proposed in the literature usually fail because such methods work correctly only for specific images. Slides from various sources may vary significantly depending on the method of smear preparation. In order to deal with these problems, an automatic segmentation procedure that integrates results of image segmentation from different methods is proposed.

3. Methodology

3.1. Preprocessing. In the process of image acquisition, the light emitted by the source passes through the glass of preparation, the specimen itself, the microscope optics, and is finally converted into digital form by a CCD sensor. Each of these steps causes distortion. Therefore, the images need to be preprocessed to obtain the best material for further analysis.

In the images from the data set presented in Section 2, there are four main types of distortion that can be removed in the image enhancement step. First, a CCD camera causes the presence of noise. Although barely visible in the original images, the noise is intensified when increasing the contrast later in the preprocessing step and may cause artifacts in the segmentation process. In order to reduce the noise, the images are filtered using a Gaussian low-pass filter (Nixon and Agudo, 2012):

\[ h_g(x, y) = \exp \left( -\frac{(x^2 + y^2)}{2\sigma^2} \right), \]

where \( \sigma = 0.85 \) was chosen experimentally, and the mask size is \( 3 \times 3 \). The filter is scaled so that the sum of all its elements equals one:

\[ h_{gn}(x, y) = \frac{h_g(x, y)}{\sum_x \sum_y h_g}. \]

Then, the images are sharpened with the following sharpening filter:

\[ h_s = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \]

The convolution is conducted for each channel of the RGB color space separately.

Another image defect requiring removal is a vignette, caused by microscope optics. In order to determine the shape and strength of the vignette, a blank slide \( I_{\text{blank}} \) was prepared as a reference. The correction is applied to the images as follows:

\[ I' = I + I(1 - I_{\text{mask}}), \]

where \( I_{\text{mask}} \) is a mask representing a decrease in brightness for each pixel and is given as

\[ I_{\text{mask}} = \frac{I_{\text{blank}}}{\max(I_{\text{blank}})}. \]

At this point the images are cropped to the size of 696 \( \times \) 568 (4 pixels from each side) to remove the frame and other artifacts that might be observed on the borders of the images. Finally, as the images have low contrast due to the lighting conditions and CCD sensor quality, histogram stretching (Nixon and Agudo, 2012) is applied.

Let \( H_{\text{input}} \) and \( L_{\text{input}} \) be the highest and lowest gray levels of image \( I \), respectively, and \( H_{\text{output}} \) and \( L_{\text{output}} \) be the highest and lowest gray levels of the range of interest, which is the maximum range possible to obtain having a given bit resolution. In the proposed approach, when the images are 24 bit RGB (8 bit per channel), and
each channel is processed separately, then $H_{\text{output}} = 255$ and $L_{\text{output}} = 0$. The input gray levels are transformed according to

$$I'_{x,y} = \frac{H_{\text{output}} - L_{\text{output}}}{H_{\text{input}} - L_{\text{input}}} (I_{x,y} - L_{\text{input}}) + L_{\text{output}}. \quad (6)$$

Using the actual highest and lowest gray levels of the input image $I$ as the parameters $H_{\text{input}}$ and $L_{\text{input}}$ not always gives satisfactory results due to outliers in the histogram. Instead, values that specify the top $H_\theta\%$ and the bottom $L_\theta\%$ of all pixel intensities can be applied. In the proposed approach, the parameters equal $H_\theta = 1\%$ and $L_\theta = 1\%$.

3.2. Nuclei segmentation. To cope with the nuclei segmentation, a two-step segmentation procedure is proposed. In the first step, a hybrid method based on adaptive thresholding, $k$-means clustering and conditional erosion is used to discover centers of nuclei. In the second step, the multi-label fast marching algorithm initialized with the discovered centers is used to extract individual nuclei.

The whole procedure starts from converting the original image $I$ to the binary image $BW$ with nuclei region highlighted. The binary image $BW$ is the result of image processing using adaptive thresholding and $k$-means clustering. Adaptive thresholding is applied to distinguish all dark objects (nuclei, cytoplasm, erythrocytes) from bright background. The image is segmented into two subsets, $G_{\text{dark}}$ and $G_{\text{bright}}$ (Sezgin and Sankur, 2003):

$$G_{\text{bright}} = \{g_{i,j}: g_{i,j} > t_{i,j}\},$$
$$G_{\text{dark}} = \{g_{i,j}: g_{i,j} \leq t_{i,j}\}, \quad (7)$$

where $g_{i,j}$ is the pixel luminance value and the threshold $t_{i,j}$ is calculated adaptively for subsequent pixels of the image using the averaging filter

$$t_{i,j} = \frac{1}{mn} \sum_{k=-n}^{n} \sum_{l=-n}^{n} g_{i+k,j+l}, \quad (8)$$

with $m$ as the size of the filter window (an odd integer) and $n = (m - 1)/2$. Unknown values outside the bounds of the image are assumed to equal the nearest image boundary value.

Another processing is necessary to distinguish nuclei from the rest of the dark objects. This task is realized using $k$-means clustering (Lloyd, 1982). In the case considered, three clusters are defined. The clusters correspond to nuclei, erythrocytes and cytoplasm. The clustering procedure is carried out in the RGB color space on the subset of pixels $G_{\text{dark}}$.

The clustering procedure of the $k$-means algorithm is based on minimizing the within-cluster sum of squares:

$$J = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{k=1}^{K} \mu_{i,j,k} D^2_{i,j,k}, \quad (9)$$

where $N_x$ and $N_y$ define the size of the analyzed image, $u_{i,j,k}$ is a function specifying whether the $(i,j)$-th pixel belongs to the $k$-th cluster, $D^2_{i,j,k}$ is the squared Euclidean distance measure,

$$D^2_{i,j,k} = (c_{i,j} - \bar{v}_k)^T (c_{i,j} - \bar{v}_k), \quad (10)$$

where $c_{i,j} \in \mathbb{R}^3$ is a vector of the coordinates of the $(i,j)$-th pixel in the RGB space and $\bar{v}_k \in \mathbb{R}^3$ is a vector of the coordinates of the $k$-th cluster center in the RGB space. The algorithm iteratively changes pixel assignments based on the distance to the nearest mean (cluster center) and updates the cluster centers to match the proper means of clusters (Hartigan and Wong, 2001). Finally, pixels are distributed into three clusters. The cluster corresponding to the nuclei is determined based on the fact that nuclei are the darkest objects in the image. Next, pixels that belong to the nuclei cluster are used to construct binary image $BW$. It marks regions in the image where the nuclei are located. At the end of image processing all objects touching the image border are removed. It can be observed that at this stage of image segmentation some nuclei are properly segmented but there is also a lot of nuclei that are stuck together (Fig. 2(b)). For this reason, further processing is necessary to separate the clustered nuclei.

A key stage of the proposed segmentation procedure is to correctly mark nuclei centers to seed the fast marching algorithm. The method is based on the concept of conditional erosion (Yang et al., 2006). The procedure assumes that the erosion is conducted as long as the size of the processed nucleus is large enough. Two masks for erosion operation are designed. They can be referred to as fine and coarse erosion structuring elements. Coarse erosion tends to preserve the actual shape but reduces the size of clustered nuclei. This can make the nucleus disappear because of huge reduction in the size. On the other hand, the fine erosion mask is less likely to make the nucleus disappear, but it will lead to loss of the original shape.

The erosion operation of the binary image $I$ by the structuring element $B$ is defined by

$$I \ominus B = \{x \in \mathbb{R}^2 \mid (B + x) \subset I\}, \quad (11)$$

where $\hat{B}$ is a reflection of set $B$. Conditional erosion is applied to binary image $BW$ obtained in the previous step of segmentation. The threshold $T_1$ for the coarse structuring element $B_c$ and the threshold $T_2$ for the fine structuring element $B_f$ are chosen experimentally ($T_1 = 350$, $T_2 = 50$). Next, nuclei are iteratively eroded using the coarse element until the size of all objects is smaller than $T_1$. Finally, erosion with the fine element is applied iteratively to the results obtained during coarse processing. Structuring elements $B_c$ and $B_f$ are designed according to the
shape of the nuclei, which is similar to an ellipse:

$$B_c = \begin{bmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 1 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix},$$  \hspace{1em} (12)

and

$$B_f = \begin{bmatrix}
0 & 1 & 0 \\
1 & 1 & 1 \\
0 & 1 & 0
\end{bmatrix}.$$  \hspace{1em} (13)

The means of objects that have survived conditional erosion become initial seeds used by fast marching to segment individual nuclei (Fig. 2(c)).

The fast marching method is a special case of the level set approach for monotonically advancing fronts. It was introduced by Sethian (1996) and can be used to extract complex shapes from 2D and 3D images. In our work it is used to split clustered nuclei. The algorithm starts with the initial front $$\Gamma_0$$. Next, the front $$\Gamma$$ evaluates with speed $$F(x,y)$$ in the normal direction, where $$F$$ is always either positive or negative. The front passes through a point $$(x,y)$$ at the time $$T(x,y)$$. Under this formulation, arrival time function $$T(x,y)$$ satisfies the Eikonal equation:

$$|\nabla T|F = 1.$$  \hspace{1em} (14)

In order to solve the equation, the gradient $$|\nabla T|$$ is estimated using the upwind entropy-satisfying scheme. By limiting our discussion to a two-dimensional grid, we must solve the following quadratic equation:

$$\frac{1}{F_{i,j}^2} = \max\left( \max(d_{i,j}^{-}\hat{T},0), -\min(d_{i,j}^{+}\hat{T},0) \right)^2 + \max\left( \max(d_{i,j}^{-}\hat{y}T,0), -\min(d_{i,j}^{+}\hat{y}T,0) \right)^2,$$  \hspace{1em} (15)

where

$$d_{i,j}^\pm\hat{T} = \frac{1}{\hat{k}}(T_{i\pm1,j} - T_{i,j}),$$

$$d_{i,j}^\pm\hat{y}T = \frac{1}{\hat{k}}(T_{i,j\pm1} - T_{i,j}),$$  \hspace{1em} (16)

and $$\hat{k}$$ is the grid step. If the quadratic equation yields more than one solution, the greatest is chosen.

The basic idea of the fast marching procedure is to propagate the front from smaller values of $$T$$ to larger ones. The algorithm constructs the narrow band around the initial front and next marches this band forward, freeze the values of existing points and brings new ones into the narrow band (Malladi and Sethian, 1996). The procedure is repeated until the narrow band is empty. The behavior of the front is driven by the speed function $$F$$. It must be designed in such a way that the front stops exactly at the boundary of the nuclei. We decided to use the speed function based on the image local gradient:

$$F = \exp\left(-\alpha|\nabla(H_\sigma * I)|\right).$$  \hspace{1em} (17)

where $$\alpha$$ is a weighting factor, $$I$$ is the original image and $$H_\sigma$$ is a Gaussian smoothing operator.

Standard fast marching is well suited to foreground-background segmentation. Nevertheless, our application must deal with multiple objects. It was realized by using multi-label fast marching (Sifakis and Tziritas, 2001; Stec, 2005). The number of labels is determined by the number of nuclei detected by conditional erosion. Fast marching is initiated by the seeds corresponding to nuclei centers. Each seed is associated with the unique label (segment). Propagation speed is the same for all labels. The algorithm maintains a single narrow band which contains trial points from all segments. Further steps are similar to the standard algorithm because a single trial point is marched forward during the single iteration. The algorithm can handle this since it can propagate fronts of any topology. A new label for the trial point is inherited from the segment that propagates at the current algorithm iteration. In order to prevent “leakages” of the nuclei segments into background and to reduce computational costs, all points classified as background by adaptive thresholding and $$k$$-means are excluded from fast marching propagation.

The final result of segmentation is recorded as matrix $$L$$ which stores labels of pixels coming from the image $$I$$. Each nucleus has its own unique integer label $$\omega_i$$ given by the fast marching procedure. The pixels labeled $$\omega_1$$ make up the first nucleus, the pixels labeled $$\omega_2$$ make up the second nucleus, and so on (Fig. 2(d)). The pixels labeled with zero are the background. Matrix $$L$$ can be directly used to compute the features of nuclei.

3.3. Feature extraction. For each isolated nucleus, 28 features are extracted. Then, for each image, the mean, median and standard deviation are determined, giving a total number of 84 features.

The features chosen reflect the observations of cytologists and can be divided into three groups. The first group is related to the size and shape of the nuclei. This is represented by the following features:

- **Area**: the actual number of pixels of the nucleus,
- **Perimeter**: the distance between each adjoining pair of pixels around the border of the nucleus,
- **Eccentricity**: the scalar that specifies the ratio of the distance between the foci of the ellipse that has the same second moments as the segmented nucleus and its major axis length,
- **Major Axis Length**: the length of the major axis of the ellipse that has the same normalized second central moments as the nucleus,
- **Minor Axis Length**: the length of the minor axis of the ellipse that has the same normalized second central moments as the nucleus.
Fig. 2. Input images, final and intermediate results of the proposed segmentation procedure: input images (a), results of nuclei region detection (b), results of conditional erosion (c), results of fast marching (d).

The second group of features is related to the distribution of nuclei in the image. Healthy tissue usually form single-layered structures while cancerous cells tends to break up, which increases the probability of encountering separated nuclei. To express this relation, we use features representing the distance to the centroid of all nuclei, and
the distance to \( k \)-nearest nuclei:

- **Distance to Centroid of All Nuclei** (dist. to all): the distance between the geometric center of the nucleus and centroid of all nuclei,

- **Distance to \( c \)-Nearest Nuclei** (dist. to \( c \)NN): the sum of distances between the geometric center of the nucleus and geometric centers of \( c \)-nearest nuclei; after conducting experiments with different values of \( c \), we decided to set this parameter to 1.

The last group of features is related to the distribution of chromatin in the nuclei. This is represented with texture features based on the Gray-Level Co-occurrence Matrix (GLCM) (Haralick et al., 1973) and the Gray-Level Run-Length Matrix (GLRLM) (Tang, 1998), as well as the mean and variance of pixel values in each RGB channel.

First four textural features are based on the GLCM. The \( N \times N \) matrix \( P \), where \( N \) is the number of gray levels, is defined over an image to be the distribution of co-occurring values of pixels at a given offset. In other words, each element of \( P \) specifies the number of times a pixel with gray-level value \( i \) occurs shifted by a given distance to a pixel with the value \( j \). Here, we calculate the mean of four GLCM features determined for offsets corresponding to \( 0^\circ, 45^\circ, 90^\circ \) and \( 135^\circ \) using eight gray-levels. In the following, \( p \) is the normalized co-occurrence matrix:

- **Contrast**: the intensity contrast between a pixel and its neighbor over the whole image:

\[
\text{contrast} = \sum_{i,j=1}^{N} |i - j| p(i,j), \quad (18)
\]

- **Correlation**: the correlation of a pixel to its neighbor over the whole image:

\[
\text{correlation} = \sum_{i,j=1}^{N} \frac{(i - \mu_i)(j - \mu_j)p(i,j)}{\sigma_i \sigma_j}, \quad (19)
\]

- **Energy** (in the literature also known as uniformity): the sum of squared elements in the GLCM:

\[
\text{energy} = \sum_{i,j=1}^{N} p(i,j)^2, \quad (20)
\]

- **Homogeneity**: the closeness of the distribution of elements in the GLCM to the GLCM diagonal:

\[
\text{homogeneity} = \sum_{i,j=1}^{N} \frac{p(i,j)}{1 + |i - j|}, \quad (21)
\]

The next eleven textural features are based on the gray-level run length matrix. The \( N \times M \) matrix \( p \), where \( N \) is the number of gray levels and \( M \) is the maximum run length, is defined for a given image as the number of runs with pixels of gray level \( i \) and run length \( j \). Similarly as in the GLCM, we compute run length matrices for \( 0^\circ, 45^\circ, 90^\circ \) and \( 135^\circ \) using eight gray-levels:

- **Short Run Emphasis (SRE)**:

\[
\text{SRE} = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{j^2}, \quad (22)
\]

- **Long Run Emphasis (LRE)**:

\[
\text{LRE} = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j)j^2, \quad (23)
\]

- **Gray-Level Nonuniformity (GLN)**:

\[
\text{GLN} = \frac{1}{n_r} \sum_{i=1}^{M} \left( \sum_{j=1}^{N} p(i,j) \right)^2, \quad (24)
\]

- **Run Length Nonuniformity (RLN)**:

\[
\text{RLN} = \frac{1}{n_r} \sum_{i=1}^{N} \left( \sum_{j=1}^{M} p(i,j) \right)^2, \quad (25)
\]

- **Run Percentage (RP)**:

\[
\text{RP} = \frac{n_r}{n_p}, \quad (26)
\]

where \( n_r \) is the total number of runs and \( n_p \) is the number of pixels in the image.

- **Low Gray-level Run Emphasis (LGRE)**:

\[
\text{LGRE} = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2}, \quad (27)
\]

- **High Gray-level Run Emphasis (HGRE)**:

\[
\text{HGRE} = \frac{1}{n_r} \sum_{i=1}^{N} \sum_{j=1}^{M} p(i,j) i^2, \quad (28)
\]

- **Short Run Low Gray-level Emphasis (SRLGE)**:

\[
\text{SRLGE} = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2 j^2}, \quad (29)
\]
3.4. Classification. For classification we use three classifiers (Bishop, 2006; Duda et al., 2001): k-Nearest Neighbor (kNN) (Cover and Hart, 1967) using $k = 5$, Decision Tree (DT) (Breiman et al., 1984), and Support Vector Machine (SVM) (Cortes and Vapnik, 1995) using a third-order polynomial kernel. The parameters were chosen experimentally (e.g., to determine the optimal $k$ for the $k$NN, we performed tests for $k = 1, 2, \ldots, 24, 25$ and chose one that gave the best result).

Classification performance was evaluated with the $n$-fold cross-validation procedure (Devijver and Kittler, 1982). There were 50 folds (the number of patients), and each fold consisted of 9 images that belong to a single patient. Two measures of the classification accuracy were defined:

- patient accuracy: the percentage ratio of successfully diagnosed cases (patients) to the total number of cases,
- image accuracy: the percentage ratio of successfully classified images to the total number of images.

All of the images were classified individually to determine image accuracy. Patient accuracy was acquired by means of majority voting for all of the classifier outcomes related to a given patient. For example, if at least 5 of 9 images were benign, then the final diagnosis was benign. Based on the majority voting results, patient accuracy was determined for each classification algorithm. Majority voting imitates a diagnostic procedure used by the pathologists who need to analyze few fragments of the slide to deliver the final diagnosis. More reliable results came from such an approach because the decision was made after a series of images had been processed rather than only a single image.

A suboptimal set of features was determined using a sequential forward selection algorithm. The procedure starts from the classifier which does not have any input variables. Then, the set of input variables is recursively expanded. The variable that most increases the accuracy of the classifier is added to the suboptimal subset. The complete procedure is repeated until no improvement can be detected. To evaluate a single set of input variables, the whole $n$-fold cross validation procedure must be carried out. The competing sets are compared in terms of the image classification accuracy. The set determined by forward selection was then used to specify the maximum classification accuracy.

4. Experimental results

The system was tested with 450 real medical images (see Section 2). The nuclei were segmented using the method described in Section 3.2. Then, for each image, 84 features were extracted as in Section 3.3. The classification accuracy was tested using three different classification algorithms and the $n$-fold cross validation technique (Devijver and Kittler, 1982). The images belonging to the same patient were never at the same time in the training and the testing set. The diagnosis for patients was determined by majority voting as described in Section 3.4.

To investigate the discriminative power of the features, image accuracy was calculated for each single feature using the $k$NN classifier. Results of this study are presented in the Table 1. For 6 features, image accuracy was over 80%, for 36 features, image accuracy was in the range of 60%–80%, and for 42 features, image accuracy was below 60%. In the second study, sequential forward selection described in Section 3.3 was employed to find the suboptimal set of features. The procedure chose 9 features for the $k$NN classifier (median of mean G value, standard deviation GLRLM SRHGE, median GLRLM SRLGE, area mean, standard deviation of variance of B value, mean of mean B value, median of dist. to all, median of minor axis length, mean GLRLM GLN), 5 features for the SVM classifier (mean GLRLM SRLGE, mean GLRLM SRHGE, mean dist. to cNN, standard deviation of GLCM...
homogeneity, median of mean R value) and 3 features for the DT classifier (mean of GLRLM SRLGE, median of GLRLM SRE, mean of eccentricity). As shown in Table 2, each classifier gave 100% classification accuracy for patients. This means that all patients were diagnosed correctly. However, we observed differences in the accuracy for individual images. The best result was obtained using $k$NN (99.33%) while the worst was obtained using the DT (91.56%).

In the previous studies, we adopted different clustering methods and adaptive thresholding to segment the nuclei (Filipczuk et al., 2011b; 2011a; Kowal et al., 2011a; 2011b). Unfortunately, all of these segmentation approaches share a major drawback. They could not separate closely spaced nuclei and this leaded to a distortion of the features that described the nuclei. In the current study, the segmentation method has been modified to deal with the mentioned problem. Visual inspection of the segmentation results show that the proposed approach copes well with closely spaced and overlapping nuclei (Fig. 2(d)). Most of the clumped and clustered nuclei are correctly separated, resulting in a reduction of a distortion of the features. Furthermore, we expanded the set of features by the adoption of texture features of the nuclei. The improvements applied resulted in image classification accuracy increased by 7%–8% compared with the authors’ former studies.

5. Conclusions

Breast cancer diagnosis using cytological images is a very difficult challenge. The content of such images is highly complex and its analysis in an automated way is difficult. However, we succeeded in developing a computer-aided diagnosis algorithm that can classify fine-needle biopsies as benign or malignant with high accuracy and provide valuable information for a medical specialist.

Experiments carried out on real medical data proved that conditional erosion is a very useful tool for detecting nuclei centers even when the nuclei are clustered. We expected that the fast marching method initiated by these centers would be able to correctly separate such difficult structures of nuclei. This was confirmed for a vast majority of cases. The method improved classification results compared with our previous works when nuclei separation step was not used.

An interesting extension of the proposed segmentation method can be application of a more sophisticated clustering method (Boryczka, 2009). Future work will also be focused at improving the system by developing more sophisticated methods for feature selection and classification (Woźniak and Krawczyk, 2012; Boryczka and Kozak, 2010). Moreover, we plan to enlarge the set of test images.

### Table 1. Classification accuracy for individual images determined for all 84 individual features (mean, median and standard deviation (STD) calculated for all 28 nuclei features, see Section 3.3). The classification was performed using the $k$NN. All values are expressed as a percentage (%).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Median</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td>63.11</td>
<td>59.78</td>
<td>58.44</td>
</tr>
<tr>
<td>perimeter</td>
<td>70.22</td>
<td>69.11</td>
<td>47.56</td>
</tr>
<tr>
<td>eccentricity</td>
<td>51.78</td>
<td>48.89</td>
<td>47.78</td>
</tr>
<tr>
<td>major axis length</td>
<td>66.44</td>
<td>64.89</td>
<td>45.78</td>
</tr>
<tr>
<td>minor axis length</td>
<td>62.67</td>
<td>61.56</td>
<td>48.89</td>
</tr>
<tr>
<td>dist. to all</td>
<td>74.89</td>
<td>72.22</td>
<td>69.11</td>
</tr>
<tr>
<td>dist. to cNN</td>
<td>75.11</td>
<td>75.11</td>
<td>70.44</td>
</tr>
<tr>
<td>GLCM contrast</td>
<td>58.00</td>
<td>60.67</td>
<td>52.89</td>
</tr>
<tr>
<td>GLCM correlation</td>
<td>60.44</td>
<td>59.78</td>
<td>61.56</td>
</tr>
<tr>
<td>GLCM homogeneity</td>
<td>54.89</td>
<td>56.89</td>
<td>60.44</td>
</tr>
<tr>
<td>GLCM energy</td>
<td>55.78</td>
<td>53.78</td>
<td>70.89</td>
</tr>
<tr>
<td>GLRLM SRE</td>
<td>68.67</td>
<td>69.11</td>
<td>76.22</td>
</tr>
<tr>
<td>GLRLM LRE</td>
<td>73.78</td>
<td>69.78</td>
<td>38.67</td>
</tr>
<tr>
<td>GLRLM GLN</td>
<td>57.78</td>
<td>61.78</td>
<td>56.00</td>
</tr>
<tr>
<td>GLRLM RLN</td>
<td>52.22</td>
<td>57.78</td>
<td>47.11</td>
</tr>
<tr>
<td>GLRLM PR</td>
<td>59.33</td>
<td>53.11</td>
<td>54.44</td>
</tr>
<tr>
<td>GLRLM LGRE</td>
<td>53.11</td>
<td>54.22</td>
<td>60.89</td>
</tr>
<tr>
<td>GLRLM HGRE</td>
<td>50.67</td>
<td>57.11</td>
<td>51.56</td>
</tr>
<tr>
<td>GLRLM SRLGE</td>
<td>84.89</td>
<td>81.78</td>
<td>83.11</td>
</tr>
<tr>
<td>GLRLM SRHGE</td>
<td>63.78</td>
<td>54.22</td>
<td>75.11</td>
</tr>
<tr>
<td>GLRLM LRLGE</td>
<td>58.67</td>
<td>58.22</td>
<td>72.89</td>
</tr>
<tr>
<td>GLRLM LRHGE</td>
<td>54.89</td>
<td>56.67</td>
<td>50.00</td>
</tr>
<tr>
<td>mean R value</td>
<td>59.78</td>
<td>63.33</td>
<td>48.67</td>
</tr>
<tr>
<td>mean G value</td>
<td>80.00</td>
<td>85.11</td>
<td>50.00</td>
</tr>
<tr>
<td>mean B value</td>
<td>79.78</td>
<td>82.00</td>
<td>55.33</td>
</tr>
<tr>
<td>variance of R value</td>
<td>53.78</td>
<td>58.22</td>
<td>48.00</td>
</tr>
<tr>
<td>variance of G value</td>
<td>62.44</td>
<td>63.11</td>
<td>58.67</td>
</tr>
<tr>
<td>variance of B value</td>
<td>70.89</td>
<td>66.44</td>
<td>79.78</td>
</tr>
</tbody>
</table>

### Table 2. Classification results for suboptimal subsets of features.

<table>
<thead>
<tr>
<th></th>
<th>$k$NN</th>
<th>DT</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy (patients)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>accuracy (images)</td>
<td>99.33%</td>
<td>91.56%</td>
<td>97.56%</td>
</tr>
</tbody>
</table>

### Acknowledgment

The authors wish to thank Dr. Roman Monczak from the Regional Hospital in Zielona Góra, Poland, for his great help and interesting discussions.

This research was partially supported by the National Science Centre in Poland.

Pawel Filipczuk is a scholar within Sub-measure 8.2.2: Regional Innovation Strategies, Measure 8.2: Transfer of knowledge, Priority VIII: Regional human resources for the economy, Human Capital Operational Programme, co-financed by the European Social Fund and state budget.
References


Marek Kowal received his M.Sc. and Ph.D. degrees in electrical engineering from the University of Zielona Góra, Poland, in 2000 and 2004, respectively. Currently, he is an assistant professor in the Institute of Control and Computation Engineering at the same university. He has published about 40 papers in refereed journal and conference papers. He is an author of one monograph and four book chapters. His current interests include medical diagnostics, image processing, fuzzy logic, as well as fuzzy and stochastic geometry. Doctor Kowal has taken part in the realization of a number of research projects sponsored by the State Committee for Scientific Research in Poland and the European Commission: INCO-Copernicus on Integration of quantitative and qualitative fault diagnosis methods within the framework of industrial application and the 5th FP EU RTN on Development and application of methods for actuator diagnosis in industrial control systems, DAMADC. He is currently involved in the realization of the research project on Diagnosis of breast cancer based on microscopic images of the material acquired by fine needle biopsy without aspiration, sponsored by the National Science Centre in Poland. He was the chair of the organizing committee of the 11th International Conference on Diagnostics of Processes and Systems, DPS2013 (Łagów Lubuski, Poland, 2013). He is also a member of the IEEE.

Paweł Filipczuk received his M.Sc. and Ph.D. degrees in computer science from the University of Zielona Góra, Poland, in 2007 and 2013, respectively. Currently, he is an assistant lecturer at the Institute of Control and Computation Engineering, University of Zielona Góra. He is an author of over 17 papers in refereed journal and conference papers. His current interests include computer-aided diagnosis, medical imaging, and pattern recognition.

Received: 1 February 2013
Revised: 26 September 2013