Cell nuclei segmentation in cytological images using fuzzy clustering and ellipse fitting algorithm

Marek Kowal*

*University of Zielona Góra,
Institute of Control and Computation Engineering
(e-mail: M.Kowal@issi.uz.zgora.pl).

Abstract: Reliable segmentation of cell nuclei from microscopic images is an important task in many medical studies. This paper presents a novel method for the segmentation of cell nuclei from microscopic images based on fuzzy clustering and ellipse fitting algorithm. It was designed specifically to segment nuclei in breast cancer FNB (Fine Needle Biopsy) cytological images. The segmentation approach takes both the color information and the spatial information into account during detecting the cell nucleus. Because of the presence of the elliptic shape function dissimilarity measure is able to differentiate the pixels with similar color but located in different regions of the image. Moreover, it is very easy to extract morphometric parameters of the cell nuclei based on parameters of detected ellipses. Simulations and experimental results are provided to demonstrate the performance and generality of the proposed method. The proposed algorithm is able to segment cell nuclei obtained from microscopy imaging with reasonable accuracy.

Keywords: image segmentation, fuzzy clustering, parameter optimization.

1. INTRODUCTION

Breast cancer is the most common cancer among women. The prognosis in breast cancer is strongly dependent on the disease development before any treatment is applied so the chance of recovery is a function of time of the detection of cancer. Modern medicine does not provide one hundred percent reliable, if possible cheap and at the same time non-invasive diagnostic methods for the diagnosis of breast pathology. As a result, in practice the important function acting in breast cancer diagnosis is the so-called triple-test, which is based on the summary of results of three medical examinations with different degrees of sensitivity and it allows to achieving high confidence of diagnosis. The triple-test includes self examination (palpation), mammography or ultrasonography imaging and fine needle biopsy (Underwood (1987)). Fine needle biopsy is collecting nucleus material directly from tumor for microscopic verification. Next, the material (collected cells) is examined using microscope in order to confirm or exclude the presence of cancerous cells. The present approach requires a deep knowledge and experience of the cytologist responsible for diagnosis. In short, some pathologists can diagnose better than others. In order to make the decision independent of the arbitrament factor, morphometric analysis can be applied. Objective analysis of microscopic images of cells has been a goal of human pathology and cytology since the middle of the 19th century. Early work in this area consisted of simple manual measurements of cell and nuclear size. Along with the development of advanced vision systems and computer science, quantitative cytopathology has become a useful method for the detection of diseases, infections as well as many other disorders (Suri et al. (2002); Hrebien et al. (2008)). In this work, we present a method that allows recognizing the malignancy and distinguish cancerous cells from the normal and benign cells. The classification of the tumor is based on morphometric examination of nuclei cells. In contrast to normal and benign nuclei, which are typically uniform in appearance, cancerous nuclei are characterized by irregular morphology that is reflected in several parameters. Morphometric measurements characterizing the shape and size have been mainly used for feature extraction. Features of the nuclei can be extracted from the image after the nuclei are correctly segmented and labeled. This work concentrates mainly on the segmentation phase because this stage of image processing is critical for correct diagnosis and treatment. Taking into account the specificity of the problem, the segmentation task is used to distinguish nuclei from the background and next label all nuclei. One of more often used methods for image segmentation is fuzzy clustering in the form of the Fuzzy C-Means algorithm (FCM). Unfortunately, the algorithm is useless in the case considered because it makes it possible to distinguish the brighter background from darker groups of cell nuclei and it is required to separate each single nuclei from the image in order to measure its morphometric parameters. Finally, the fuzzy c-means with shape algorithm has been adopted to tackle the mentioned problem (Leung et al. (2004); Wang et al. (2007)). The original version of the FCMS algorithm was prepared to determine two clusters only (background and demanded objects). The work proposes a modified iterative algorithm.
in order to overcome the mentioned problem and allows determining all nuclei in the image as separate objects.

The paper is divided into three sections. Section 1 gives an overview of breast cancer diagnosis and fine needle biopsy imaging techniques. Section 2 describes the idea of the FCMS algorithm. Section 3 deals with modifications applied to the original algorithm. Section 4 shows the experimental results obtained using the proposed approach. The last part of the work includes a summary, conclusions and references.

2. FUZZY C-MEANS INCORPORATING ELLIPTIC FUNCTION

The idea of the segmentation of images using a clustering algorithm usually boils down to a search for natural clusters of pixels represented in color space. With this approach, it is possible to find objects characterized by similar colors, and by applying the fuzzy clustering algorithm the method is more robust against disturbances and image inaccuracy (Babuska (1998); Rutkowska and Zadeh (2000); Szczepaniak et al. (2000)). Taking into account the fine needle biopsy images of nuclei, the task of segmentation reduces to searching separate objects (cell nuclei) in order to determine their morphometric parameters. Unfortunately, the application of fuzzy clustering in the form of the FCM algorithm allows us to separate two interesting objects only: the background and all cell nuclei viewed as one big object. In order to solve the problem, a modified FCM algorithm is used, which was originally developed in the work (Leung et al. (2004); Wang et al. (2004)) and used in lip segmentation from the face image.

The FCMS algorithm, similarly to the original FCM algorithm, is based on finding the local minimum of the nonlinear cost function using the Picard iteration through first-order conditions for stationary points (Bezek (1981); Babuska (1998)). In the case of the FCMS algorithm, the cost function is defined by the following expression:

\[ J(U, V, \theta) = \sum_{x=1}^{X} \sum_{y=1}^{Y} \sum_{k=1}^{C} \mu_{x,y,k}^{m}(f_c(x,y,v_k) + \alpha f_d(x,y,\theta,n_k)) \],  

where the matrix \( U \in \mathbb{R}^{X \times Y \times C} \) contains the membership degrees of pixels to the defined clusters, \( V = [v_1, v_2, \ldots, v_C] \), \( v_i \in \mathbb{R}^{C \times 4} \) is a matrix which defines the centers of the clusters, \( \theta \) is a vector of parameters describing the preferred shape of the objects searched in spatial domain, \( X \) and \( Y \) defines the size of the analyzed image, \( f_c \) is a function used to determine the distance between the data points and cluster centers, \( f_d \) is a function to evaluate the belong rate of the pixel to the ellipse associated with the tested cluster, \( \mu_{x,y,k} \) is the membership of the \((x,y)\)-th pixel in the fuzzy cluster \( k \), \( m \in (1,\infty) \) is the fuzziness of the clustering procedure, \( c_{x,y} \) is a vector of \((x,y)\)-th pixel parameters which describes its color, \( v_k \in \mathbb{R}^4 \) is a vector of the coordinates of the \( k \)-th cluster center, \( n_k \) is an exponent of elliptic function associated with the \( k \)-th cluster, \( \alpha \) is a weighting parameter.

The function \( f_c \) defines dissimilarity measures based on the distance between the data points and cluster centers:

\[ f_c(x,y,v_k) = \parallel x - y - v_k \parallel^2 = (x - y - v_k)^T A (x - y - v_k) \]

The matrix \( A \) which occurs in the expression (2) is used to tune the shape and orientation of the clusters in space. In the simplest approach, the matrix \( A \) is unitary, thus the distance measure \( f_c(x,y,v_k) \) is an Euclidean norm. In this case, the study metric is defined as a Euclidean distance which measures color dissimilarity in RGB or HSV color space. The function \( f_d \) incorporates the shape function in the objective function and measures the spatial distance. As the nuclei are more like an ellipse, it was decided to chose the shape function \( f_d \) as elliptic function described by the following expression:

\[ f_d(x,y,\theta,n_k) = \left( \frac{(x - x_e) \cos \phi + (y - y_e) \sin \phi)^2}{w^2} + \frac{(x - x_e) \cos \phi - (y - y_e) \sin \phi)^2}{h^2} \right)^{n_k} \],

where \( \theta = [x_e,y_e,w,h,\phi] \) denotes the set of parameters that describes the elliptic function, \((x_e,y_e)\) is the center of the ellipse, \( w \) and \( h \) are respectively the semi-major and semi-minor axis, and \( \phi \) is inclination angle, exponent \( n_k \) ensures a small function values within the ellipse and large values for the pixels outside the ellipse. Figure 1 shows values of the sample elliptic function \( f_d \) for two clusters. Taking into account that each cluster has identical parameters \( \theta \) but with a different exponent \( n_k \), it is possible to determine two clusters only. The first one represents the background and the second one represents the cell nuclei found in the image.

![Fig. 1. Sample elliptic function for two clusters](image-url)
The idea of the proposed fuzzy clustering is to minimize the objective function (1):
\[
(U^*, V^*, \theta^*) = \arg \min_{(U, V, \theta)} J(U, V, \theta).
\] (4)

The minimum of the objective function is calculated by an iterative algorithm using the following expressions for the updating matrices \( U \) and \( V \):
\[
\mu_{x,y,k} = \left[ \frac{C}{j=1} \left( f_c(c_{x,y}, v_j) + \alpha f_d(x, y, \theta, n_k) \right)^{\frac{1}{1-\gamma}} \right]^{-1},
\] (5)
\[
v_k = \sum_{x=1}^{X} \sum_{y=1}^{Y} \mu_{x,y,k} I_{x,y}.
\] (6)

where \( I_{x,y} \in \mathbb{R}^q \) represents a single pixel. The parameters of the ellipse \( \theta \) are updated in each step of the FCMS algorithm by the gradient descent algorithm using the following general expression:
\[
\theta_i(t+1) = \theta_i(t) - \gamma \frac{\partial J(U, V, \theta)}{\partial \theta_i},
\] (7)

where
\[
\frac{\partial J(U, V, \theta)}{\partial \theta_i} = \alpha \sum_{x=1}^{X} \sum_{y=1}^{Y} \sum_{k=1}^{C} \mu_{x,y,k} f_d(x, y, \theta, n_k) \frac{\partial f_d(x, y, \theta, n_k)}{\partial \theta_i}.
\] (8)

Equations (5), (6) and (7) are iterated as a Picard iteration to obtain the segmentation of the image. Detailed expressions for updating the parameters of the ellipse can be found in the paper (Leung et al. (2004)).

Figure 2 shows a sample image with two cell nuclei and a corresponding objective function calculated for two parameters \( x_c \) and \( y_c \) (for simplicity, the rest of the ellipse parameters are frozen). Two local minima present in Fig. 2 represent two cell nuclei present in the original image. Hence, the gradient descent method can be applied to find the local minima and thus the centers of cell nuclei.

It can be noticed that the objective function is continuous in \((U, V, \theta)\) and \( I_{x,y} \) is bounded in \( \mathbb{R}^q \). Moreover, the second partial derivative of the objective function with respect to \((U, V)\) is a positive diagonal matrix for any feasible \( \theta \). Following the proof of convergence for FCMS in Bezdek (1981) and FCMS in Leung et al. (2004) respectively, the objective function value of the \( i \)-th Picard solution \((U^i, V^i)\) is less than that of \((U^{i-1}, V^{i-1})\).
\[
J(U^i, V^{i-1}, \theta^{-1}) < J(U^{i-1}, V^{i-1}, \theta^{-1}),
\] (9)
and
\[
J(U^i, V^i, \theta^{-1}) < J(U^i, V^{i-1}, \theta^{-1}).
\] (10)

Using the gradient descent method for solving \( \theta_i \), the objective function \((U^{i-1}, V^{i-1})\) is decreasing until \( \theta_i \) falls to a local minimum or reaches the boundary of the \( \theta \)-vector space. Taking this into account, the following inequality can be formulated:
\[
J(U^i, V^i, \theta^i) < J(U^i, V^i, \theta^{i-1}).
\] (11)

Therefore, the Picard iteration is shown to give a local minimum solution. Unfortunately, in the presented approach the update of the parameter vector \( \theta \) is computationally very expensive and the solution is extremely dependent upon the initial values of the calculated parameters. Figure 5 shows the result of fuzzy segmentation applied to find cell nuclei with random initialization of ellipse parameters. It can be observed that the optimization procedure very often leads to the detection of spurious cell nuclei corresponding to the local minimum of the objective function.

The next section tackles the presented problems and finally solves them by applying some modifications to the segmentation procedure presented in this section.

3. MODIFIED FCMS ALGORITHM

The fuzzy segmentation method presented in the previous section is able to differentiate the objects with different colors, but what is more significant in the case considered it is also designed to differentiate objects with the same color but located in different regions of the image. Such a feature of the algorithm is very useful in the segmentation of cell nuclei. Unfortunately, the original FCMS algorithm (Leung et al. (2004); Wang et al. (2004)) has a few drawbacks and limitations when applied to the segmentation of cell nuclei. The main problem arises from the fact that the algorithm is able to find only two clusters, the first one corresponding to the object with an elliptic shape and the second one corresponding to the rest of the image, while it is required to discover as much as possible of cell nuclei in order to make reliable diagnosis based on cell nuclei parameters.

In Section 2 it was mentioned that the FCMS algorithm is computationally very expensive due to gradient descent
optimization applied to solve the parameters of the ellipse in each iteration of the whole FCMS procedure. Moreover, the results obtained using such a procedure are very problematic because the objects found are usually not cell nuclei but rather groups of cell nuclei or parts of cell nuclei. Taking into account this fact, it was decided to solve the ellipse parameters optimization problem only once after the FCM algorithm has preprocessed the original image by minimizing (12). So, the modified segmentation procedure can be viewed as minimizing the objective function (12) and afterwards minimizing the objective function (13) with respect to $\theta$.

$$J_1(U, V) = \sum_{x=1}^{X} \sum_{y=1}^{Y} \sum_{k=1}^{C} \left[ \mu_{x,y,k}^m f_c(c_{x,y}, v_k) \right], \quad (12)$$

$$J_2(\theta) = \sum_{x=1}^{X} \sum_{y=1}^{Y} \sum_{k=1}^{C} \left[ \mu_{x,y,k}^m f_d(x, y, \theta, n_k) \right]. \quad (13)$$

It was observed during the experiments that the standard FCMS algorithm usually requires about 15 iterations to detect separate cell nuclei and the whole segmentation procedure must be repeated for each cell nuclei present in the image, so the introduced modification dramatically decreased the number of computations required to segment the image.

Unfortunately, the developed method does not solve all of the mentioned problems because, similarly like standard FCMS, it often gets stuck in an inappropriate minimum (Fig. 5). This problem arises from the fact that the parameters of the initial ellipse were generated randomly and the gradient descent method converges to the nearest local minimum. To overcome the problem, it was decided to limit the initial values of ellipses using knowledge about a typical size of the cell nuclei and the multi-start gradient method (Fig. 6). The method allows us to reduce false segmentation of cell nuclei; however, finally the best results was obtained using the Hough like transform to initialize the ellipse. Such an approach proposes systematic generation of a series of ellipses arranged in the form of grids covering the whole image (Fig. 3). Next, the objective function is calculated for each ellipse and the ellipse with the smallest value of the objective function (13) becomes a starting point for the gradient descent tuning procedure. Of course, such a procedure was time consuming, so in order to reduce the complexity, the grids of ellipses was sparse and the parameters $w$ and $h$ were limited by the knowledge about a typical size of the cell nuclei, the parameter $\phi$ was constant to simplify the exploration. The proposed initialization greatly improved the quality of the segmentation and accelerated the convergence of the gradient method. In order to reduce the number of steps required to reach the minimum by the gradient method even more, an adaptive learning rate was applied to control the level of parameter updates.

The scheme of the final version of the algorithm is presented in Fig. 4. It consists of three steps. In the first one, the original image is preprocessed by the FCM method to find the color of cell nuclei and the background. Next, the parameters of ellipses are initialized using the developed approach. Finally, the gradient method is used to tune the parameters of ellipses in the following iterations. The ellipse found a single algorithm iteration represents a single cell nucleus. In order to facilitate the searching procedure, cell nuclei already detected and described by the corresponding ellipse are erased from the original image by filling in the ellipses using the background color. The stop criteria for the gradient descent algorithm are defined in the form of a minimal value of objective function change and a maximum number of tuning steps. The stop criteria
Table 1. Dataset

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples (malignant)</td>
<td>50 (25)</td>
</tr>
<tr>
<td>Number of images</td>
<td>500</td>
</tr>
<tr>
<td>Format of files</td>
<td>BMP</td>
</tr>
<tr>
<td>Resolution</td>
<td>704x578</td>
</tr>
<tr>
<td>Number of colors</td>
<td>256</td>
</tr>
</tbody>
</table>

for global algorithm iterations are defined as the maximum number of ellipses found and the minimum objective function value (13). Choosing right values for stop criteria requires some kind of experience from the user. The next section presents some sample results of segmentation using the developed approach.

4. EXPERIMENTAL RESULTS

Testing existing and newly developed algorithms requires to have databases at disposal, on which tests and benchmarks can be realized, especially in the domain of image analysis, where in many problems a domain knowledge needs to be taken into account. In our study we decided to design a new data set that could be applied to a completely automatic process of image analysis. The segmentation of cell nuclei was carried out on cytological material obtained by FNB. Biopsy without aspiration was performed under the control of ultrasonography with a needle of a diameter of 0.5 mm. Smears from the material were fixed in spray fixative (Cellfix by Shandon) and dyed with hematoxylin and eosin (h+e). The time between the preparation of smears and their preserving in fixative never exceeded three seconds. The smears were derived from 25 FNB of benign and 25 of malignant lesions collected from 50 patients of the out-patient clinic ONKOMED in Zielona Góra . All cancers were histologically confirmed and all patients with benign disease were either biopsied or followed for a year. The image for digital analysis was generated by a SONY CCD IRIS color video camera mounted at top of AXIOPHOT microscope. The slides were projected into a camera with 10× and 160× objective and a 2.5× ocular. One image was generated for 100× enlargement and nine for 400× enlargement.

Most of the criteria of malignancy are seen in the nuclei of the cells. Therefore, it is essential to isolate the nuclei from the rest of the image. It can be observed that malignant cells in contrast to normal and benign cells are characterized by irregular morphology that is reflected in several morphometric parameters. Morphometric measurements characterizing the shape and size have been mainly used for feature extraction. The extracted features are: size, circularity, perimeter, compactness, lengths of the axis of the ellipse circumscribing the nuclei and the eccentricity of the ellipse circumscribing the nuclei. Details about these features can be found in (Wolberg et al. (2004); Marciniak et al. (2005); Jeleń et al. (2008); Obuchowicz et al. (2008)). However, from previous research in the subject it is known that there are big differences in size between benign and malignant cases, and shape factors do not have good discriminative properties (Marciniak et al. (2005)). Hence the proposed segmentation approach seems to be very promising in this case due to simplicity in calculating crucial features. In order to test the effectiveness of the proposed segmentation method, a set of

Fig. 5. Segmentation using standard FCMS

Fig. 6. Segmentation using FCM clustering and gradient descent with random initialization

Fig. 7. Segmentation using FCM clustering and gradient descent with grid initialization

50 manually binary-segmented images was prepared. The result of automatic segmentation was compared to that of manual segmentation using the error measure based on the Hamming distance between manually segmented image and automatic segmented image. Since the developed segmentation method produces as a result of segmentation a set of ellipses, it was needed to convert ellipses which represent cells to a binary image by filling in the ellipses. The results of the comparison are summarized in Table 2, where EF stand for ellipse fitting phase. Following error measures were used to demonstrate the performance of segmentation:

\[
E_{\text{ave}} = \sum_{i=1}^{N} \frac{e_i}{n_iN}, \quad (14)
\]

\[
E_{\text{min}} = \min_{i=1...N} \frac{e_i}{n_i}, \quad (15)
\]

\[
E_{\text{max}} = \max_{i=1...N} \frac{e_i}{n_i}, \quad (16)
\]

where \(e_i\) is a Hamming distance between the \(i\)-th manually segmented binary image and the \(i\)-th automatically segmented binary image, \(n_i\) is the number of cell nuclei in the \(i\)-th image and \(N\) is the number of test images. It must be noticed that although FCM and KM clustering achieved better segmentation results with respect to defined criteria
than FCMS and two stage FCMS with random initialization, their usage is strongly limited due to the fact that they are not able to determine separated cell nuclei. In order to compare the segmentation results visually, some segmented parts of images are shown in Figs. 5, 6 and 7.

The software to realize the segmentation initially was prepared in the Matlab environment; unfortunately, time consuming computations make this software applicable to parts of the original images only. Hence the final version of the software was prepared in the Java environment using the Java Advanced Imaging (JAI) library. In order to accelerate the computations, the segmentation algorithm was written as a JAI operator and was prepared to work in a multi-threaded environment. All experiments were done using a PC class machine with a quad core processor 3.2GHz.

<table>
<thead>
<tr>
<th>Segmentation method</th>
<th>( E_{ave} )</th>
<th>( E_{max} )</th>
<th>( E_{min} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCMS (RGB)</td>
<td>132</td>
<td>25670</td>
<td>29</td>
</tr>
<tr>
<td>FCM (RGB), EF with random init.</td>
<td>127</td>
<td>22198</td>
<td>32</td>
</tr>
<tr>
<td>FCM (HSV), EF with grid init.</td>
<td>78</td>
<td>17011</td>
<td>15</td>
</tr>
<tr>
<td>FCM (RGB), EF with grid init.</td>
<td>81</td>
<td>16210</td>
<td>14</td>
</tr>
<tr>
<td>FCM (RGB) clustering</td>
<td>102</td>
<td>14129</td>
<td>23</td>
</tr>
<tr>
<td>K-mean (RGB) clustering</td>
<td>125</td>
<td>16509</td>
<td>28</td>
</tr>
</tbody>
</table>

5. CONCLUSION

The work presents a cell nuclei segmentation method based on fuzzy-c means clustering with a shape function. The results achieved in the experiments seem to be very promising. The main advantage of the method is that the results of the segmentation can be directly used to calculate the morphometric parameters of the cells. Modifications proposed in the work to the original FCMS algorithm allow decreasing significantly the cost of computations. This achievement is very important due to the fact that nowadays colonoscopes are able to produce images with the resolution of 50000x50000 or even higher. Of course, the presented approach is not able to segment all cases properly so there are a few challenges for the near future. The first one concerns the segmentation of overlapped cell nuclei. The second one regards the adaptation of FCMS clustering to deal with more complex shapes. This will allow calculating parameters of cell nuclei more precisely, and possibly the algorithm can gain more general applications due to the ability to detect objects with different shapes (Gong et al. (2004); Zoller and Varvarigou (2007); Chatzis and Buhmann (2008)). However, it must be pointed out that the present accuracy of segmentation is so high that the future research will concentrate on the feature extraction and classification stages. The recognition rate in malignancy classification will be a true test for the quality and effectiveness of the developed segmentation method (Zhang (1996)).

REFERENCES


