Lymphocyte Segmentation using the Transferable Belief Model

Costas Panagiotakis[1], Emmanuel Ramasso[2], and Georgios Tziritas[1]

[1] Department of Computer Science, University of Crete, P.O. Box 2208, Heraklion, GREECE, {cpanag,tziritas}@csd.uoc.gr
[2] FEMTO-ST Institute, UMR CNRS 6174 - UFC / ENSMM / UTBM, Automatic Control and Micro-Mechatronic Systems Department, FRANCE, emmanuel.ramasso@femto-st.fr

Abstract. In the context of several pathologies, the presence of lymphocytes has been correlated with disease outcome. The ability to automatically detect *lymphocyte nuclei* on histopathology imagery could potentially result in the development of an image based prognostic tool. In this paper we present a method based on the estimation of a mixture of Gaussians for determining the probability distribution of the principal image component. Then, a post-processing stage eliminates regions, whose shape is not similar to the *nuclei* searched. Finally, a Transferable Belief Model is used to detect the *lymphocyte nuclei*, and a shape based algorithm possibly splits them under an equal area and an eccentricity constraint principle.

1 Introduction

Recently, there is an increasing activity on analysing histopathological images, as a potential prognostic tool for cancer patients. One important step for the diagnosis is the cell segmentation. Demir and Yener [1] review the different approaches classified in two categories: region-based and boundary-based methods. Lymphocyte segmentation in histopathology images is complicated by the similarity in appearance between *lymphocyte* and *cancer nuclei* in the image [2]. In [2], a computer-aided diagnosis (CADx) scheme is proposed to automatically detect and grade the extent of lymphocytic infiltration in digitized HER2+ BC histopathology. Lymphocytes are automatically detected by a combination of region growing and Markov random field algorithms using the luminance channel in Lab color space. Finally, a support vector machine classifier is used to discriminate samples with high and low lymphocytic infiltration. In [3], lymphocytes are automatically detected via a segmentation scheme comprising a Bayesian classifier and template matching, using the Saturation color channel in HSV color space.

In [4], a segmentation scheme, Expectation Maximization driven Geodesic Active Contour with Overlap Resolution (EMaGACOR), is proposed for automatically detecting and segmenting lymphocytes on HER2+ Breast Cancer histopathology images. EMaGACOR utilizes the Expectation-Maximization (EM)

algorithm for automatically initializing a geodesic active contour and includes a scheme for resolving overlapping structures. EMaGACOR was evaluated on a total of 100 HER2+ breast biopsy histology images and was found to have a detection sensitivity of over 86% and a positive predictive value (PPV) of over 64%.

Our method addresses the problem of lymphocyte detection and should be considered as a region-based approach. The first step of our method consists of a likelihood classification based on the estimation of the parameters of a mixture of Gaussians. A post-processing step eliminates regions with size or shape that differ greatly from a typical shape of *lymphocyte nuclei*. For the remaining regions the following features are extracted: mean value, variance, eccentricity and size. A Transferable Belief Model is then trained and used in order to detect the *lymphocyte nuclei*. Finally, a shape based algorithm possibly splits the detected regions under an equal area and an eccentricity constraint principle.

The organisation of the paper is as follows: Section 2 describes the segmentation stage with the estimation of a mixture of Gaussians and the shape-based detection of candidate *leymphocyte nuclei*; Section 3 presents the Transferable Belief Model used and the results of training based on the ground-truth; in Section 4 is presented our technique for solving possible overlaps. Then, the results on the ICPR contest data set are given in Section 5.

2 Segmentation

There are three possible classes corresponding to *stroma, cancer nuclei* and *lymphocyte nuclei*. We admit Gaussian distributions for the three classes and use the EM algorithm for estimating the parameters of the model. We observe that the three colour channels are strongly correlated. Therefore we start by applying principal component analysis (PCA) in order to select only one image component. Let us note $x(s)$ this component at a site s of the image grid. Let $p(x)$ be the probability density function for the principal image component. According to the mixture of Gaussians model we have:

$$p(x) = \sum_{k=1}^{3} \frac{P_k}{\sigma_k \sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} = \sum_{k=1}^{3} P_k p_k(x|\mu_k, \sigma_k^2). \quad (1)$$

The unknown parameters are the *a priori* probabilities (P_k), the mean (μ_k) and the variance (σ_k^2) values.

At first, the Max-Lloyd algorithm is used for obtaining initial parameter values. The empirical probability density function is used for the estimation. Let N denotes the number of image pixels. At i-th iteration of the EM algorithm we have:

– E-step: calculate the posterior probabilities

$$P^{(i+1)}(k|x, \theta^{(i)}) = \frac{P_k^{(i)} e^{-\frac{(x-\mu_k^{(i)})^2}{2\sigma_k^{2(i)}}}}{\sqrt{2\pi}\sigma_k^{(i)} p^{(i)}(x)}, \quad (2)$$

where θ is the set of all the unknown parameters.

– M-step: estimate the prior probabilities, the mean and the variance values as follows

$$P_k^{(i+1)} = \frac{1}{N} \sum_{s \in G} P^{(i+1)}(k|x(s), \theta^{(i)}) \tag{3}$$

$$\mu_k^{(i+1)} = \frac{1}{N P_k^{(i+1)}} \sum_{s \in G} P^{(i+1)}(k|x(s), \theta^{(i)}) x(s) \tag{4}$$

$$\sigma_k^{2(i+1)} = \frac{1}{N P_k^{(i+1)}} \sum_{s \in G} P^{(i+1)}(k|x(s), \theta^{(i)}) (x(s) - \mu_k^{(i)})^2 \tag{5}$$

The above steps are implemented using the empirical probability density for limiting the computational time. A stopping threshold of 10^{-6} is given on the relative gain per iteration for the log likelihood value.

Having the estimation of the probability density functions for the three classes the image sites are classified according to the maximum likelihood principle. Therefore, for classifying the site s to class k, the likelihood $p_k(x(s)|\mu_k, \sigma_k^2)$ is maximized.

Then, a post-processing stage follows on the regions detected as candidate *lymphocyte nuclei*, which being darker are identified by the mean value. Three region parameters are measured: the area, the eccentricity and the solidity. The area of region r (denoted A_r) is given by the number of pixels that belong to region r. Very small regions are eliminated.

The eccentricity of region r (denoted E_r) is defined by the ratio between the two principal axes of the best fitting ellipse, measuring how thin and long a region is. It holds that $E_r \geq 1$. The eccentricity can be defined by the three second order moments $m_r(1,1), m_r(2,0)$ and $m_r(0,2)$. Let (c_{rx}, c_{ry}) denote the centroid of region r (given by the set of sites O_r, (s_x, s_y) being the coordinates of a point).

$$c_{rx} = \frac{1}{A_r} \sum_{s \in O_r} s_x \tag{6}$$

$$c_{ry} = \frac{1}{A_r} \sum_{s \in O_r} s_y \tag{7}$$

$$m_r(p,q) = \sum_{s \in O_r} (s_x - c_{rx})^p (s_y - c_{ry})^q \tag{8}$$

$$E_r = \sqrt{\frac{m_r(2,0) + m_r(0,2) + \sqrt{(m_r(2,0) - m_r(0,2))^2 + 4m_r^2(1,1)}}{m_r(2,0) + m_r(0,2) - \sqrt{(m_r(2,0) - m_r(0,2))^2 + 4m_r^2(1,1)}}} \tag{9}$$

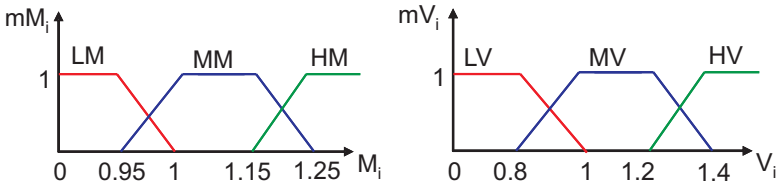
The eccentricity criterion is intended to filter line segments.

The solidity criterion measures the proportion of the pixels in the convex hull of the region that are also in the region. Therefore, it is relevant to the region shape. In our implementation a value of 2/3 is required for accepting a region as *lymphocyte nucleus* candidate.

3 Transferable Belief Model

Image and shape features are computed for each candidate region. The mean value M_i and the variance V_i of the image of a candidate region i are extracted. In order to be independent from scaling and variability in appearance, the mean and the variance of each region are normalized by division with the corresponding median values obtained on the set of all regions.

The two extracted features are combined within the *Transferable Belief Model* (TBM) framework [5] [6] in order to perform *lymphocyte nuclei* detection. The TBM is an alternative to probability measure for knowledge modelling and the main advantage and power of the TBM is the capacity to explicitly model doubt and conflict. TBM has been successfully applied on object detection and tracking problems [7] combined with shape and motion based features.



(a) (b)

Fig. 1. From numerical features to belief. **(a)** Mean value. LM, MM and HM correspond to low, medium and high values of the normalized mean of image intensity, respectively. **(b)** Variance. LV, MV and HV correspond to low, medium and high values of the normalized variance of image intensity, respectively.

The mean value and the variance can be adequately converted into beliefs (symbolic representation). This is the first step of the TBM framework. We have proposed the numeric-to-symbolic conversion presented in Fig. 1, where L is used for low value, M for medium values and H for high values. Let us note $f_k(m)$ and $g_k(v)$ the two belief functions, where $k = 1, 2, 3$ corresponds respectively to low, medium and high values. Using symbolic representation, the *lymphocyte nuclei* detection can be performed based on appropriate table rules (see Table 1). The values of Table 1 (values of $T(k, l)$) have been estimated using the ground truth images of the ICPR 2010 contest, by estimating the probability of *lymphocyte nuclei* detection for each belief pair.

Having estimated the table of rules, we compute the the plausibility B_i of each candidate *lymphocyte nucleus* region i as follows:

$$B_i = \sum_{k=1}^{3} \sum_{l=1}^{3} f_k(M_i) g_l(V_i) T(k, l) \tag{10}$$

	LV	MV	HV
LM	0.999	0.988	0.958
MM	0.565	0.669	0.701
HM	0.052	0.1	0.032

Table 1. Table rules providing $T(k, l)$ used in Equation (10).

A region i will be detected as *lymphocyte nucleus*, if $B_i > 0.55$. We have selected the threshold of 0.55, since it gives the highest accuracy results on the ground truth data set.

4 Region Splitting

Having detected the *lymphocyte nuclei* based on appearance features, we have to resolve possible overlaps using shape features. Finally, the area (A_i) and the eccentricity (E_i) [7] are used in the decision of splitting a detected region to more than one regions plausibly corresponding to *lymphocyte nuclei*.

The area and the eccentricity are normalized with report to their respective median values. According to the feature A_i, the region i can be splitted into N_i regions, where $N_i \in \{1, \ldots, \lceil A_i \rceil\}$. We split a region i into N_i possible sub-regions selecting the more appropriate splitting as described hereafter.

The proposed algorithm splits the region i into N_i equal area regions minimizing the maximum eccentricity of the resulting sub-regions j, $j \in \{1, \ldots, N_i\}$, since the *lymphocyte nuclei* are circular-like regions. A circular-like region has minimum eccentricity, close to one. Similar to the minimization of maximum error on polygonal approximation problem using equal errors criterion [8], the problem of minimizing the maximum eccentricity can be sub-optimally solved under the equal area criterion and the above eccentricity constraint. We have implemented this criterion using the following algorithm. The pseudo-code of the Region Splitting to N_i sub-regions is given in Algorithm 1.

- Initially, we sequentially select N_i seed-points $p_j, j \in \{1, \ldots, N_i\}$ of region i from which N_i parallel region growing algorithms start. The seeds should follow the next constraint so that the growing algorithms start from the farthest sub-regions: the minimum distance between all pairs of these points should be maximized.
- The optimal algorithm that solves this problem has $O(\binom{R_i}{N_i})$ computation cost, where R_i denotes the number of pixels of region i. We have used the next approximate algorithm that sub-optimally solves this problem in $O(R_i^2)$ based on the optimal solution for two regions. p_1 and p_2 are given as the two farthest points of region i (optimal solution for two regions) (lines 1-11 of Algorithm 1). The next points $p_j, j \in \{3, \ldots, N_i\}$ are sequentially computed by getting the point p of region i that maximizes the minimum of distances from p to p_{j-1}, p_{j-2},..., p_1 (lines 12-27 of Algorithm 1).

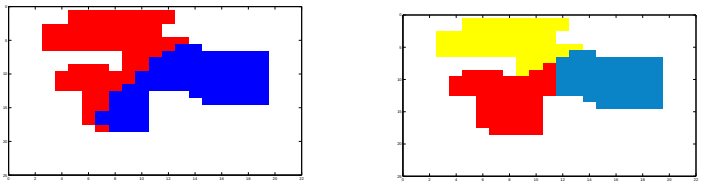
– Then N_i parallel growing algorithms start from seeds $p_j,\ j \in \{1,\ldots,N_i\}$ (lines 28-30, 31-36 of Algorithm 1). In each step, the growing algorithm $j,\ j \in \{1,\ldots,N_i\}$ adds the most close point to p_j from the set of non-visiting boundary points of sub-region j that minimizes eccentricity of sub-region j, yielding equal area regions that uniformly grow with a circular-like shape having minimal eccentricity (line 33 of Algorithm 1).

Finally, we select splitting to N_i regions, where N_i maximizes the following criterion:

$$C(N_i) = \begin{cases} \dfrac{B_i}{\sqrt{\max(A_i, \dfrac{1}{A_i}) \cdot E_i}}, & N_i = 1 \\ \dfrac{(1 - b(N_i)) \max_{j \in \{1,\ldots,N_i\}} B_{i,j}}{\sqrt{\max(\bar{A}, \dfrac{1}{\bar{A}}) \cdot \bar{E}}}, & N_i > 1 \end{cases} \tag{11}$$

where $\bar{A}$ and $\bar{E}$ denote the mean area and the mean eccentricity of the N_i split regions. $b(N_i)$ denotes the percentage of boundary pixels between the resulting sub-regions (intrinsic boundary pixels) of splitting. $B_{i,j}$ denotes the plausibility of *lymphocyte nuclei* sub-region for the sub-region j of region i estimated by TBM framework. This criterion is maximized when the mean area and mean eccentricity is close to one (that corresponds to most appropriate shape for *lymphocyte nucleus* region) and the maximum probability of *lymphocyte nucleus* sub-region is high.

Fig. 2 illustrates an example of region splitting algorithm execution for $N_i = 2$ and $N_i = 3$. According to ground truth, the algorithm successfully gives three partitions, since for $N_i = 3$ the proposed criterion was maximized, $C(1) = 0.47, C(2) = 0.24, C(3) = 0.53, C(4) = 0.35$.



(a) (b)

Fig. 2. An example of Region Splitting into **(a)** $N_i = 2$. **(b)** and $N_i = 3$ sub-regions.

input : Region O_i. Number of sub-regions N_i that O_i will be split.
output: The N_i sub-regions R_i^j, $j \in \{1, ..., N_i\}$.

```
1  d_max = 0
2  foreach (x_1, y_1) ∈ O_i do
3      foreach (x_2, y_2) ∈ O_i do
4          d = (x_1 − x_2)^2 + (y_1 − y_2)^2
5          if d > d_max then
6              d_max = d
7              p_1 = (x_1, y_1)
8              p_2 = (x_2, y_2)
9          end
10     end
11 end
12 for j = 3 to N_i do
13     d_max = 0
14     foreach (x_1, y_1) ∈ O_i do
15         d_min = ∞
16         for n = 1 to j − 1 do
17             d = (p_n.x − x_1)^2 + (p_n.y − y_1)^2
18             if d < d_min then
19                 d_min = d
20             end
21         end
22         if d_min > d_max then
23             d_max = d_min
24             p_j = (x_1, y_1)
25         end
26     end
27 end
28 for j = 1 to N_i do
29     R_i^j = {p_j}
30 end
31 repeat
32     for j = 1 to N_i do
33         p̂_j = getNextPoint(j, R_i, p_j, O_i)
34         R_i^j = R_i^j ∪ {p̂_j}
35     end
36 until ∀j ∈ {1, ..., N_i} ⇒ p̂_j = ∅
```

Algorithm 1: Region Splitting Algorithm.

5 Experimental Results

We have tested our method on the data of the Pattern Recognition in Histopathological Images contest (ICPR 2010). Fig. 3 illustrates results of the proposed scheme for image *im8.tif* of the data set. Figs. 3(a) and 3(b) illustrate the original image and the principal image component, respectively. Fig. 3(c) illustrates final results of the method with ground truth. Red boundaries correspond to candidate regions that are detected as *lymphocyte nuclei* regions (see Section 3). Blue boundaries correspond to candidate regions that are not detected as *lymphocyte nuclei* regions (see Section 3). Green and white squares are the centroids of real *lymphocyte nuclei* and detected regions, respectively. Fig. 3(d) illustrates final detection of the proposed method (white regions). The region that belongs in [75,85] $\times$ [45,55] bound box has been successfully splitted into two sub-regions. Similarly with Fig. 3(c), Fig. 4 depicts the final results of the method with ground truth for the rest images of dataset.

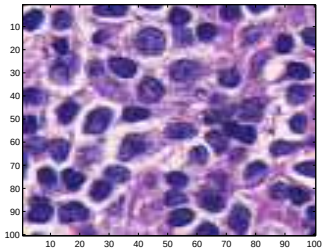
Image	*Sensitivity*	*PPV*
im1	0.968	0.815
im2	0.961	0.714
im3	0.900	0.720
im4	0.950	0.791
im5	0.965	0.933
im6	0.944	0.756
im7	0.928	0.928
im8	0.883	0.926
im9	0.941	0.592
im14	0.952	0.869

Table 2. Sensitivity and PPV.

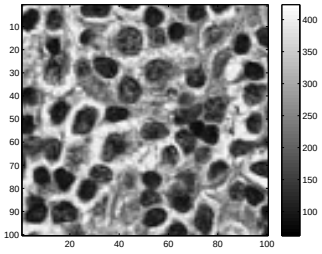
Table 2 depicts the Sensitivity and the PPV for each image of the tested data set. According to this table, Sensitivity and PPV take values in range $[0.928, 0.968]$ and $[0.714, 0.926]$, respectively. The mean values of Sensitivity and PPV are 0.938 and 0.807, respectively.

6 Conclusion

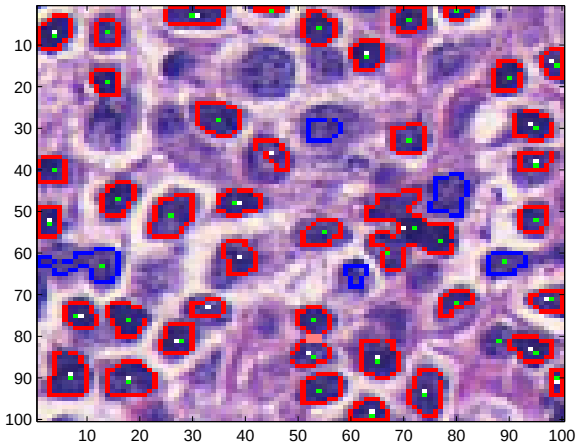
We have proposed an appearance and shape based method for automatic detection of *lymphocyte nuclei* on histopathology images. We have used a mixture of Gaussians for determining the probability distribution of the principal image component and the TBM framework with a region splitting method to detect and split the *lymphocyte nuclei* regions. The proposed algorithm gives high accuracy results on the whole data set: Sensitivity of 0.938 and PPV of 0.807.



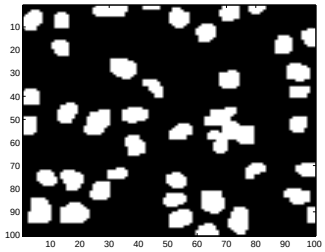
(a)



(b)

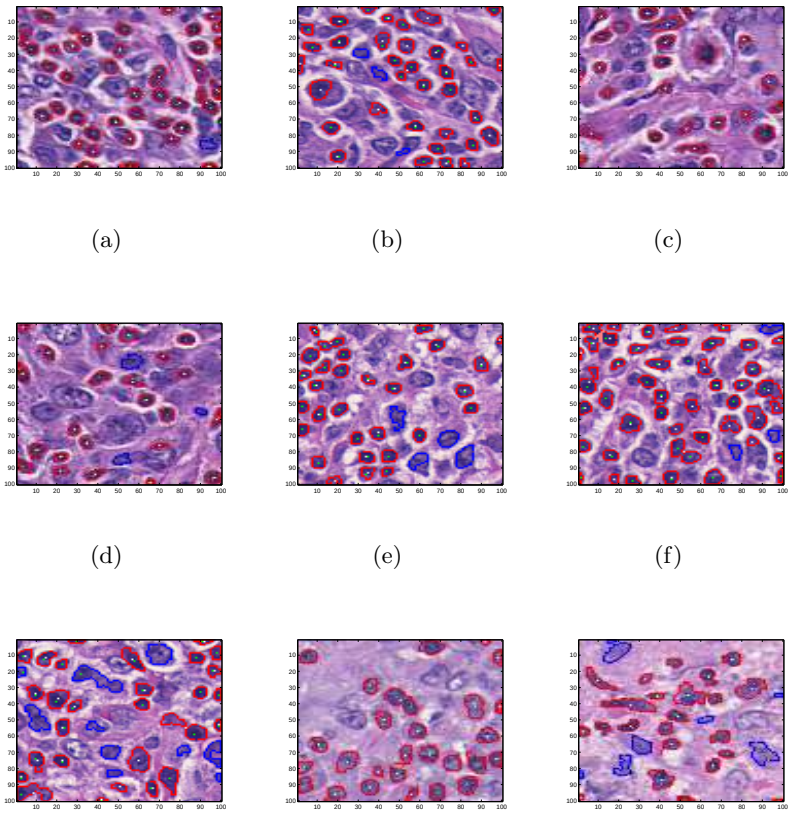


(c)



(d)

Fig. 3. (a) The original image. **(b)** The one channel image after PCA. **(c)** The final detection with ground truth. **(d)** The final detected regions.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 4. The final detection with ground truth.

Acknowledgements

The work of Costas Panagiotakis has been supported by postdoctoral scholarship (2009-10) from the Greek State Scholarships Foundation (I.K.Y.).

References

1. Demir, C., Yener, B.: Automated cancer diagnosis based on histopathological images: a systematic survey. Technical Report 05-09, Rensselaer Polytechnique Institute (2005)
2. Basavanhally, A., Ganesan, S., Agner, S., Monaco, J., Feldman, M., Tomaszewski, J., Bhanot, G., Madabhushi, A.: Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. IEEE Transactions on Biomedical Engineering **57**(3) (2010) 642–653
3. Basavanhally, A., Agner, S., Alexe, G., , Ganesan, G.B.S., Madabhushi, A.: Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade breast cancer histology. In: Workshop on Microscopic Image Analysis with Applications in Biology (in conjunction with MICCAI). (2008)
4. Fatakdawala, H., Basavanhally, A., Xu, J., Bhanot, G., Ganesan, S., Feldman, M., Tomaszewski, J., Madabhushi, A.: Expectation maximization driven geodesic active contour with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on breast cancer histopathology. IEEE Transactions on Biomedical Engineering (2010 (to appear))
5. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligence **66**(2) (1994) 191–234
6. Smets, P.: Decision making in the TBM: the necessity of the pignistic transformation. Int. Jour. of Approximate Reasoning **38** (2005) 133–147
7. Panagiotakis, C., Ramasso, E., Tziritas, G., Rombaut, M., Pellerin, D.: Shape-based individual/group detection for sport videos categorization. Intern. J. Pattern Recognition Artificial Intelligence **22**(6) (2008) 1187–1213
8. Panagiotakis, C., Tziritas, G.: Any dimension polygonal approximation based on equal errors principle. Pattern Recogn. Lett. **28**(5) (2007) 582–591