A Classification Scheme for Lymphocyte Segmentation in H & E Stained Histology Images

Manohar Kuse, Tanuj Sharma, and Sudhir Gupta

The LNM Institute of Information Technology, Jaipur, INDIA
kusemanohar.08@lnmiit.ac.in, tanuj.08@lnmiit.ac.in, sudhir@lnmiit.ac.in

Abstract. A technique for automating the detection of lymphocytes in histopathological images is presented. The proposed system takes Hematoxylin and Eosin (H&E) stained digital color images as input to identify lymphocytes. The process involves segmentation of cells from extracellular matrix, feature extraction, classification and overlap resolution. Extracellular matrix segmentation is a two step process carried out on the HSV-equivalent of the image, using mean shift based clustering for color approximation followed by thresholding in the HSV space. Texture features extracted from the cells are used to train a SVM classifier that is used to classify lymphocytes and non-lymphocytes. A contour based overlap resolution technique is used to resolve overlapping lymphocytes.

Keywords: Lymphocytes, Classification, Contour Overlap Resolution

1 Introduction

A lymphocyte is a type of blood cell in the immune system. Lymphocyte count is carried out to help diagnose many ailments. The infiltration of lymphocyte has been correlated with the disease outcome in cases of breast and ovarian cancer, leukemia, acquired immuno deficiency syndrome, viral infection, etc [10]. The ability to automatically detect and quantify extent of lymphocyte infiltration on histopathology imagery could potentially result in the development of a computer assisted diagnosis tool for Her2+ and ovarian cancer [9].

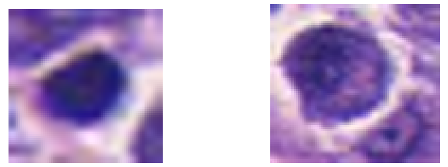
A study showed that the Lymphocytic Infiltration is relevant prognostic indicators and might be used as markers for an appropriate treatment strategy in patients with stage I carcinomas [13]. Another study claims to find strong correlations between the infiltration of lymphocytes and occurence of cancer[5].

Early detection of breast cancer is the key for its prognosis. Mammography has been one of the most reliable method of detection of breast cancers. However, enormous sizes of mammogram data had made it is difficult to manually detect breast cancer [2]. Qualitative pathological examination of the images leads to inexact classification of the cells and is subject to observer variation and variability based on the spatial focus of observation rendering the derived high level information subjective. Computer assisted diagnosis can provide objective description of the cells and assist pathologists for finding disorders associated with lymphocyte count.

Visual inspection of the histology slides does not allow one to distinguish between lymphocyte nuclei and cancer nuclei (see figure 1). Other challenges associated with automation of lymphocyte detection are the ability of the method to accommodate variability in staining procedures, differing scales of image digitization, varying illumination conditions and high occurrence of overlapping objects.

This paper describes a clinically relevant classification scheme of Hematoxylin and Eosin (H&E) stained histology slides to detect lymphocytes. The scheme is based on automated image processing, supervised learning of texture features and contour based overlap resolution. This work was done as part of a contest titled "Pattern Recognition in Histopathological images" held during International Conference on Pattern Recognition, 2010. There were a total of 10 images comprising lymphocytic infiltration that were H&E stained and digitized at 20 X resolution. 6 images were used as the training set and the other 4 were used as testing images for the results obtained in this paper. The images also came with expert annotations of representative lymphocytes. The expert annotations provided the approximate locations of centers of the lymphocytes (see figure 2(b)), and a few boundary anotations were also provided to get an idea of the shapes of lymphocytes.

Figure 2(a) shows one such histology image. Figure 2(b) and 2(c) shows the annotated centers and boundaries respectively, provided by the organizers. While the annotation of lymphocyte centers was complete, only five lymphocyte boundary annotations were provided per image.



(a) Lymphocyte by Ground Truth

(b) Non-Lymphocyte by Ground Truth

Fig. 1. Visual inspection of the histology slides does not allow one to distinguish between lymphocyte nuclei and cancer nuclei.

2 Related Work

Various techniques have been proposed to detect lymphocytes based on color, texture and shape features. Hybrid segmentation methods have been used to

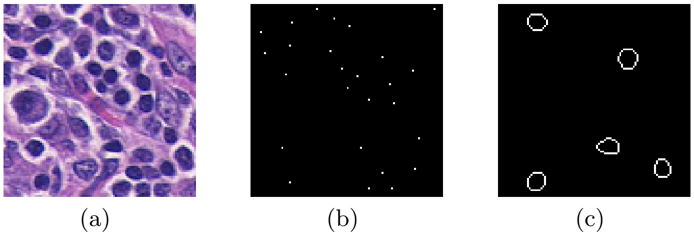


Fig. 2. Contest Dataset

detect nuclei from images of histology slides stained under different conditions [12, 16].

The watershed transformation is one of the most powerful tools for segmenting images [6] but the problem with watershed segmentation is that noisy and textured images have many minima, most of them being irrelevant for segmentation. Using the watershed on a gray tone image without any preparation leads to a strong over segmentation. The best solution to this problem consists in initially determining markers for each region of interest, including the background of the image. This makes it semi automated with subjectivity creeping in because of the choice of markers.

Active contour based models for lymphocyte segmentation have also been proposed [7], but the choice of seed points affects its segmentation performance. Bikhet et al [1] have used hierarchical thresholding to localize white blood cells, followed by extraction of gray level and morphological features to train a supervised classifier. Thresholding works well on a given set of images but fails with variability in the image set. Ongun et al [14] have used morphological preprocessing to segment the cells followed by fuzzy patch labeling.

3 Proposed Classification Scheme

The main stages of the proposed classification scheme are: 1) Extracellular matrix (ECM) segmentation, 2) Morphological pre processing, 3) Contour based overlap resolution, 4) Feature extraction, 5) Classification using a trained SVM classifier. MATLAB was used for prototyping of the scheme designed for this contest. Figure 3 shows the overview diagram of the proposed classification scheme.

3.1 ECM Segmentation

The H&E stain dyes DNA-rich cell nuclei blue and collagen-rich extracellular matrix (ECM) pink, allowing differentiation of cell from the surrounding ECM based on color [19]. Two steps are involved in the ECM segmentation. 1) Mean Shift Clustering, 2) HSV Based Thresholding.

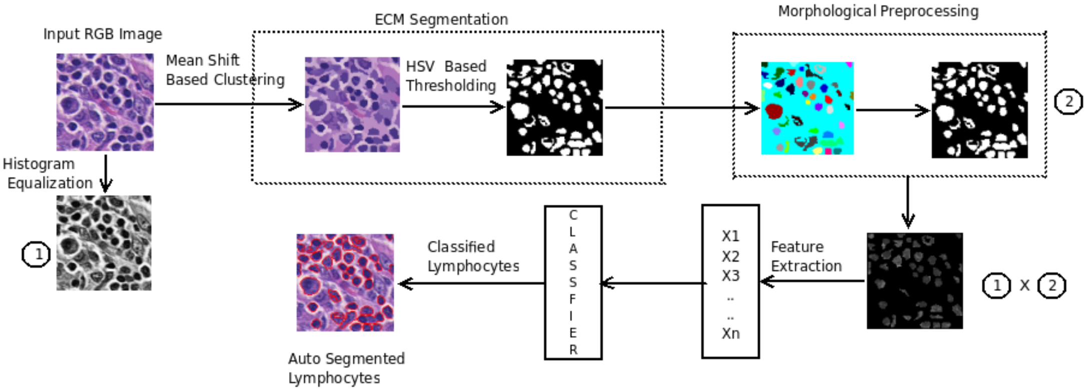
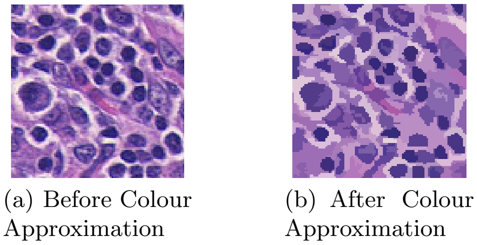


Fig. 3. Overview Diagram

Mean Shift based Clustering for Color Approximation Mean Shift Clustering is a non parametric clustering technique based on density estimation for analysis of complex feature space. Dense regions in feature space correspond to local maxima of the probability density function, i.e to the modes of the unknown density [4, 8]. Clustering was used to approximate the colors present (see figure 4) in the image to reduce computational efforts.

For example there are 4061 distinct colors present in figure 4(a). After mean shift based clustering, the number of distinct colors reduced to 172.

As an unintended consequence, it also lead to some structures being represented by similar colors which could then be easily segmented using a thresholding in HSV space.

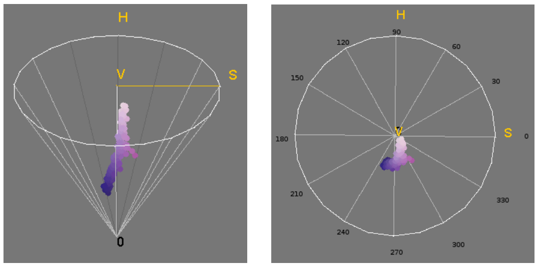

(a) Before Colour Approximation

(b) After Colour Approximation

Fig. 4. Colour Approximation using Mean Shift Clustering

HSV Based Thresholding The HSV color space corresponds closely to the human perception of color and it has been proven more accurate and effective in distinguishing colored objects. The values of the thresholding to separate pink hue from blue hue, were obtained by 3D visualization of the distribution of these colors (as shown in figure 5) using an open source software ImageJ [15]. Extracellular matrix was segmented from the cells using equation 1. Where M represents the binary mask which is being formed after thresholding.

$$M(i,j) = \begin{array}{ll} 1\ , & \text{if } 0.6667 \ \leq \ \text{hue(i , j)} \ \leq \ 0.7292 \\ 0\ , & \text{otherwise} \end{array} \tag{1}$$



(a) (b)

Fig. 5. 3D Visualization of HSV Colour Space

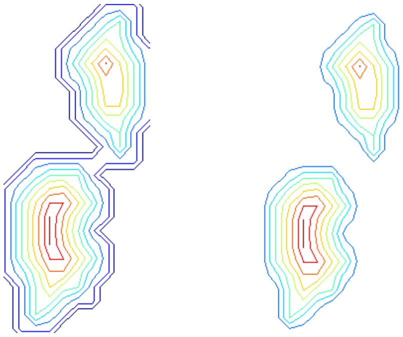
3.2 Morphological Pre-Processing

Connected components analysis (CCA) labels the the blobs in a binary image, as per its connectivity. The labels thus formed were used to iterate through each of the blobs thus formed, to extract the blob features. Overlap resolution is applied to blobs which satisfy some threshold on area and perimeter as discussed in section 3.3.

3.3 Contour Based Overlap Resolution

A novel contribution of this paper is in resolving cell overlaps. The importance of resolving overlaps in lymphocyte detection and grading is discussed in [7]. Overlaping of lymphocytes, sometimes makes it difficult to segment them.

Here we have used a contour based heuristic for revolving the overlap among the lymphocytes. Contours are defined by those pixels that are at an equal distance from the detected cell boundary. Further, those closed contours which cover an area that approximates to the area of an average lymphocyte are retained while ignoring other contours. Figure 6 shows an example to illustrate the overlap resolution.



(a) (b)

Fig. 6. Overlap Resolution

3.4 Feature Extraction

The mask obtained from the previous steps represented lymphocytes as digital number 1 and other areas as digital number 0. This mask was multiplied with the histogram equalized grayscale image of the RGB image shown in figure 2(a). Histogram equalization was performed to normalize varying illumination conditions. Eighteen texture features were extracted for every detected cell region [17, 18, 3]. These are – Autocorrelation, Contrast, Correlation, Cluster Prominence, Cluster Shade, Dissimilarity, Energy, Entropy, Homogeneity, Maximum probability, Variance, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, Information measure of correlation, Normalized inverse difference moment [17, 18, 3]. These features were derived from the gray level cooccurence matrix for four values of offset and four values of direction. Average of these eight values was used as feature value in classification.

3.5 Supervised Classification

Supervised classification was performed to classify the cells into two classes – lymphocytes and non-lymphocytes. For training the classifier, the labels for every feature pattern were obtained from the annotated dataset and a training dataset was constructed that consisted of 80 patterns for lymphocytes and 98 patterns for non lymphocytes. A support vector machine classifier was trained using this training dataset [11].

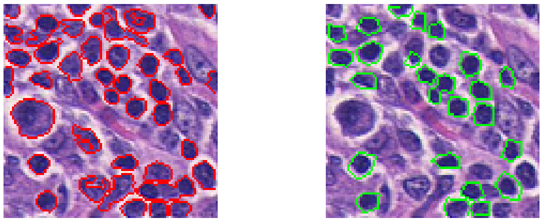
4 Results

The classification scheme described in section 3 was applied on 4 testing images that had 94 lymphocytes and 74 non lymphocytes as per expert annotation. A correct detection of lymphocyte in the confusion matrix tabulated in table 1 meant that a lymphocyte centre marked by the expert existed in the region classified as lymphocyte by the proposed scheme.

Table 1. Confusion Matrix

		Ground Truth	
		Lymphocytes	Non-Lymphocytes
Classifier	Lymphocytes	161	25
	Non-Lymphocytes	55	133

It can be observed from the confusion matrix that the proposed classification scheme is able to achieve a classification accuracy of 78% at a false positive rate of 14.7%. Figure 7(a) shows the lymphocytes detected by the proposed classification scheme as cells that are delineated with a red boundary. Figure 7(b) shows the lymphocytes annotated by the expert that were delineated with the help of given lymphocyte centers. Visual inspection of the results shows that there is good agreement between the derived results and the ground truth. The results sent by us were also evaluated by the organizers using two region based measures and two boundary based measures.



(a) Classifier Result

(b) Expert Annotated

Fig. 7. Comparison between Classifier Results and Ground Truth

The region based measures are defined as follow
1) Dice coefficient $DICE = \frac{2\times|A(S)\cap A(G)|}{|A(S)|+|A(G)|}$
2) Sensitivity $SN = \frac{|A(S)\cap A(G)|}{|A(G)|}$

The boundary based performance measures are defined as follow
1) Hausdorff distance

$$HD = max_w[\ min_x||c_w - c_x||\ ](c_w \in S, c_x \in G)$$

2) Mean absolute distance

$$MAD = \frac{\sum_{w=1}^{M} ||c_w - c_x||}{M}$$

Where C is the total number of pixels in the image and $|s|$ represents cardinality of any set s. $A(s)$ and $A(G)$ is the area of the close boundary of segmentation results and manual delineation. For boundary based measures S and G are closed boundaries of segmentation results and manual delineations. M is the number of pixels on the closed boundaries of segmentation results.

Table 2. Performance Comparison

Group	DICE	SN	HD	MAD
1	0.73	0.57	4.58	0.77
2	0.74	0.58	3.63	0.65
3	0.37	0.23	21.95	9.14
4	0.83	0.71	3.73	0.41
5	0.74	0.59	3.51	0.62

Table 2 shows the results summarized by the organizers using the above mentioned metrics. The results of group 2 correspond to the results obtained from the work mentioned in this paper. DICE coefficient is a measure of similarity of images. Our method gives 74% means that, the actual result is 74% similar to the output provided by our method. Our sensitivity is 58% means that 58% of the positives are correctly identified. It can be observed that DICE coefficient is only 0.9 less than the best reported result. There is a scope for improvement in sensitivity by the introduction of newer features related to shape and color.

5 Conclusions

We have developed a classification scheme for automatically detecting lymphocytes from H&E stained histopathology slides. However, the proposed scheme needs extensive testing on different images that are truly representative of the various scenarios in the real world. Such a dataset will also help to build a good knowledge base for supervised classification of images. Without such an extensive evaluation, a prognosis tool for lymphocyte count related disorders cannot be developed especially when the risk associated with misclassification is high.

6 Future Work

As of now, the system does not require user interaction or parameter tuning and produces classification results that are better than most methods used in the contest. The size of the lymphocytes can be determined automatically for use with overlap resolution given the scale at which the image was acquired.

Classification results are largely based on the training of the classifier and thus there is a scope of using incremental learning to keep the knowledge base updated. A dimensionality reduction exercise can help find those features that aid in classification. Further, use of relatively higher resolution images than those used in this contest can lead to a better quantification of the texture features and it is our belief that this will further increase the ability of the classifier to distinguish between lymphocytes and non lymphocytes.

References

1. S. F. Bikhet, A. M. Darwish, H. A. Tolba, and S. I. Shaheen. Segmentation and classification of white blood cells. *IEEE International Conference on Acostics, Speech and Signal Processing, ICASSP*, pages 550–553, 1992.
2. H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du. Approaches for automated detection and classification of masses in mammograms. *Pattern Recogn.*, 39(4):646–668, 2006.
3. D A. Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Can. J. Remote Sensing*, 28:45–62, 2002.
4. Dorin Comaniciu. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 58:71–96, 2002.
5. Vanessa Deschoolmeester, Marc Baay, Eric Van Marck, Joost Weyler, Peter Vermeulen, Filip Lardon, and Jan Vermorken. Tumor infiltrating lymphocytes: an intriguing player in the survival of colorectal cancer patients. *BMC Immunology*, 11(1):19, 2010.
6. R. Doughri, S. M'hiri, K.B. Romdhane, F. Ghorbel, and S. Essafi. Segmentation and classification of breast cancer cells in hostological images. *Information and Communication Technology*, 2006.
7. Hussain Fatakdawala, Ajay Basavanhally, Jun Xu, Gyan Bhanot, Shridar Ganesan, Michael Feldman, John Tomaszewski, and Anant Madabhushi. Expectation maximization driven geodesic active contour with overlap resolution : Application to lymphocyte segmentation on breast cancer histopathology. *International Conference on Bioinformatics and Bioengineering*, 2009.
8. Fukunaga, Keinosuke, and Larry D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory (IEEE)*, 21(1):32–40, 1975.
9. PR in HIMA. http://bmi.osu.edu/cialab/ICPR_contest.
10. A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun. Cancer statistics 2008. *CA Cancer J Clin*, 58:71–96, 2008.
11. T. Joachims. Making large-scale svm learning practical. advances in kernel methods - support vector learning. *MIT-Press*, 1999.
12. L. Latson, B. Sebek, and K.A. Powell. Automated cell nuclear segmentation in colour images of he stained breast biopsy. *Anal Quant Cytol Histol*, 25(321-331), 2003.
13. Lorena Losi, Giovanni Ponti, Carmela Di Gregorio, Massimiliano Marino, Giuseppina Rossi, Monica Pedroni, Piero Benatti, Luca Roncucci, and Maurizio Ponz de Leon. Prognostic significance of histological features and biological parameters in stage i (pt1 and pt2) colorectal adenocarcinoma. *Pathology - Research and Practice*, 202(9):663 – 670, 2006.
14. Guclu Ongun, Ugur Halici, Kemal Leblebiciogl, and Volkan Atalay. Feature extraction and classification of blood cells for an automated differential blood count system. *IEEE IJCNN*, (2461-2466), 2001.
15. Rasband W. S. Imagej. *U.S. National Laboratory of Health, Bethesda, Maryland, USA*, 1997-2005.
16. F Schnorrenberg, C Pattichis, K Kyriacou, and C Schizas. Computer-aided detection of breast cancer nuclei. *IEEE Transaction on Information Technology in Bio-medicine*, pages 128–140, 1997.
17. K. Shanmugam, R. M. Haralick, and I. Dinstein. Textural features of image classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3:6, Nov. 1973.

18. L. Soh and C. Tsatsoulis. Texture analysis of sar sea ice imagery using gray level co-occurrence matrices. *IEEE Transactions on Geoscience and Remote Sensing*, 37:2, Mar. 1999.
19. Scarff R. W. and Torloni H. Histological typing of breast tumors. international classification of tumors. *World Health Organization*, pages 13–20, 1968.