TEXTURE CLASSIFICATION USING NONLINEAR COLOR QUANTIZATION: APPLICATION TO HISTOPATHOLOGICAL IMAGE ANALYSIS

Olcay Sertel[1,2], Jun Kong[1,2], Gerard Lozanski[3], Arwa Shana'ah[3],
Umit Catalyurek[1,2], Joel Saltz[2], Metin Gurcan[2]

[1]Department of Electrical and Computer Engineering [2]Department of Biomedical Informatics
[3]Department of Pathology, The Ohio State University, Columbus, OH 43210, USA

ABSTRACT

In this paper, a novel color texture classification approach is introduced and applied to computer-assisted grading of follicular lymphoma from whole-slide tissue samples. The digitized tissue samples of follicular lymphoma were classified into histological grades under a statistical framework. The proposed method classifies the image either into low or high grades based on the amount of cytological components. To further discriminate the lower grades into low and mid grades, we proposed a novel color texture analysis approach. This approach modifies the gray level co-occurrence matrix method by using a non-linear color quantization with self-organizing feature maps (SOFMs). This is particularly useful for the analysis of H&E stained pathological images whose dynamic color range is considerably limited. Experimental results on real follicular lymphoma images demonstrate that the proposed approach outperforms the gray level based texture analysis.

***Index Terms*— color texture analysis, self-organizing feature maps, computer-aided diagnosis**

1. INTRODUCTION

Follicular Lymphoma (FL) is a one of the most common non-Hodgkin B cell lymphomas in the western world with a highly variable clinical course. Patients with indolent FL do not benefit from early therapy. In fact, early chemotherapy for them may cause more harms than benefits; therefore should be avoided. On the other hand, FL patients with aggressive disease should receive appropriate therapy as soon as possible to increase their chance of remission and to prolong their lives. These important clinical decisions are currently guided by histological grading of the tumor. As recommended by the World Health Organization (WHO), histological grading of FL is based on the number of large malignant cells, namely centroblasts (CB), per standard 40× high power microscopic field (HPF) of 0.159 mm^2. In this method, centroblasts are manually counted in ten random neoplastic follicles and the average of CB/HPF [1] is reported. In this grading system, grade I corresponds to 5 or less CB/HPF, grade II to 6-15 CB/HPF and grade III to 15 or more CB/HPF. Although it is very important in clinical practice, this manual method suffers from well-documented inter- and intra-reader variability. For instance, in a multi-site study, the agreement among experts for the various grades of follicular lymphoma varied between 61% and 73% [2]. Moreover, for practical reasons, pathologists typically count CBs only in ten neoplastic follicles, leading to sampling bias. Possible consequences of over or under grading of FL include inappropriate timing and type of therapy with serious clinical consequences for patients. Therefore, we are developing a computer-assisted system for automated grading of FL with a better consistency.

Parallel to the developments in digital scanning technologies, research on histopathological image analysis is becoming more and more active. Recently, several image analysis approaches have been proposed for different types of cancers such as prostate [3], neuroblastoma [4] and colon cancers [5]. All of these studies exploit the texture information and construct the subsequent analysis over a statistical classification framework. However, most of the texture models are derived from gray-level images. The color information is incorporated after separating the color from the illumination from which the texture information is extracted and combined with the color information.

Among many texture models, gray-level co-occurrence method introduced by Haralick *et al.* is one of the most widely used texture analysis approach [6]. However, this approach is limited to gray-level images. Arvis *et al.* proposed a multi-spectral method and a uniform quantization method to incorporate the color and the texture information in the co-occurrence matrix framework [7]. The basic idea behind the multi-spectral method is to use the cross-correlation between channels to construct several co-occurrence matrices. In the latter approach, instead of using the gray-levels, color images are quantized to extract several color classes and the co-occurrence matrix uses the label of the classes for its computation. These studies conclude that the color texture approach improves the performance remarkably.

In this paper, we propose to use the color texture information from H&E-stained images for the automated

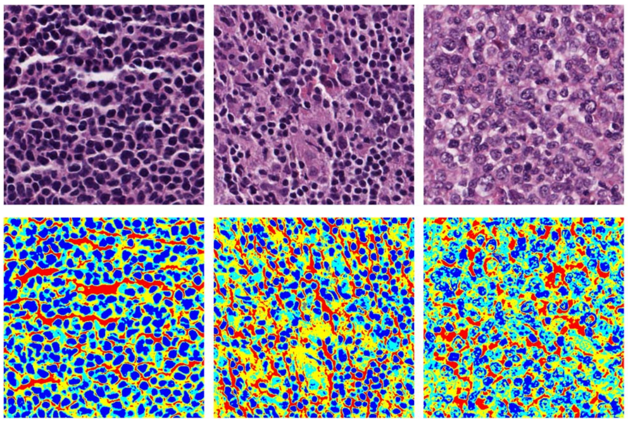

Fig.1. Sample H&E-stained images and their segmentation results associated with grade I, grade II and grade III, respectively from left to right. Blue corresponds to nuclei, cyan to cytoplasm, yellow to extracellular material and red and gray to background and red blood cells, respectively.

grading of FL. Due to the staining process, the color spectrums of these images have considerably limited dynamic ranges. Therefore, we proposed a non-linear quantization using self-organizing feature maps *(*SOFMs) and used the quantized images to construct the co-occurrence matrix. The performance of the proposed approach is compared to that of the gray-level and uniform color quantization approaches.

2. FEATURE EXTRACTION AND CLASSIFICATION

The inputs to our system are H&E-stained tissue samples digitized using a Scope XT digitizer (Aperio, San Diago, CA) at 40× magnification. In our study, we used 17 whole-slide images that cover the typical range of tissue variations for each grade. A consensus of hematopathologists identified six of these as grade I, eight as grade II and three as grade III. We asked three experienced pathologists to extract ten regions of interests from each whole-slide sample that are equivalent to one microscopic HPF, resulting in a data set of 510 images.

2.1. Image Segmentation

Initially, we focused on differentiating the low and intermediate grades (I and II) from high grades (III) of FL. The lower grades of FL samples contain larger proportion of centrocytes characterized by small size, cleaved to irregular nuclei with dispersed chromatin, inconspicuous nucleoli and scant cytoplasm, while most of the high grade cases contain a higher population of centroblasts characterized by larger size, round nuclei with vesicular chromatin and accentuated nuclear membrane.

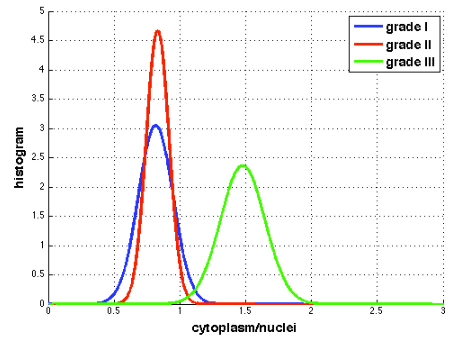

Fig.2. Distribution of the number of samples based on the ratio of the amount of cytoplasm regions to the amount of nuclei regions over the FL cases associated with three different grades.

To capture this information, we first partitioned the images into distinct cytological components using an unsupervised segmentation. Typically in H&E stained FL images, there are five major components, i.e., nuclei, cytoplasm, background, red blood cells (RBCs) and extra-cellular material and each of them is expressed in hues of different colors. At the segmentation step, we used the La*b* color space, where the difference between two colors are perceptually uniform; therefore the Euclidean distance can be used as a measure [8]. RBCs and background regions show relatively uniform patterns; thus they are segmented by thresholding the intensity values in the RGB color space. The rest of the structures were segmented by the K-means clustering algorithm with the number of clusters as three to represent nuclei, cytoplasm and extra-cellular material. Figure 1 shows sample H&E stained FL images and their segmentation results.

Cells in the high-grade samples have open chromatin with more membranes as compared to the compact cells in lower-grades with scant cytoplasm; hence the amount of different cytological components in the tissue varies considerably between low and high grades. We used the ratio of the amount of cytoplasm to the amount of nuclei, γ, as a discriminative feature to differentiate low- and mid-grade FL cases form high-grades as shown in Eq.1

$$\gamma = \frac{\chi}{\nu} \tag{1}$$

where χ, ν are number of cytoplasm and nuclei pixels, respectively. Figure 2 shows the distribution of samples based on this ratio, which was found to be very discriminative in separating high grades (III) from lower grades (I and II).

2.2. Color Texture Analysis Using SOFM

To further discriminate the low- and mid-grade FL samples, we introduced a novel way of extracting color texture information. We proposed to apply a non-linear quantization using the self-organizing feature maps *(*SOFM) and extract the gray level co-occurrence features from the quantized color labels. For our application of FL grading, the proposed approach is more intuitive than the uniform quantization-based methods introduced in [7] in several ways. First, H&E images have considerably limited color spectrum, i.e., there are very few dominant colors (hues of blue, purple and pink) as shown the sample images in Figure 1. Therefore, instead of using a uniform quantization, we can enhance the image by using a non-linear quantization. Second, uniformly quantizing the color channels provides no natural order of the colors, which is essential for computing the co-occurrence matrix.

Co-occurrence matrices are one of the most commonly used methods in texture classification [6]. Let $\{I(x,y), 0 \le x \le N-1, 0 \le y \le M-1\}$ denote a grayscale image, where G is the number of gray levels and $N \times M$ is the image size, we can construct a $G \times G$ gray-level co-occurrence matrix, P_d, that gives the joint distribution of pair of pixels with a given spatial relationship determined by the displacement vector $d(dx, dy)$ as follows:

$$P_d(i,j) = |\{((r,s),(t,v)) : I(r,s) = i, I(t,v) = j\}| \quad (2)$$

where $(r,s), (t,v) \in N \times M$, $(t,v) = (r+dx, y+dy)$, and |.| is the cardinality of a set. Subsequently, a set of features is extracted from this matrix to describe the texture of the image. In our study, we extracted five features namely, homogeneity, energy, contrast, correlation and entropy [6].

In order to construct the co-occurrence matrix, we first applied a non-linear quantization using SOFM algorithm, which is an artificial neural network that is trained using unsupervised learning [9]. Suppose, C is a set of all colors in the RGB color space denoted by $C = \{c_i, i = 1,2,...,L\}$ where L is the number of colors in the image. Color quantization is the process of choosing K colors from C to construct the quantized color space, $\overline{C}$, where $\overline{C} = \{\overline{c}_j, j = 1,2,...,K\}$.

SOFM is an iterative vector quantization method based on competitive learning, where the quantization vector is initialized and tuned iteratively. There are three basic steps involved in the algorithm at each iteration:

- **Initialization:** Assign random values for the initial weight vector, $\overline{C}(0)$.
- **Similarity Matching:** For every color C_i, in the image, compute the best matching neuron $q(\overline{C})$, using the Euclidean distance as follows:

$$q(\overline{C}) = \arg\min_j \left|C_i - \overline{C}_j\right|, j = 1,2,...,K \quad (3)$$

- **Updating:** The weight vectors of all the neurons are updated as follows:

$$\overline{C}_j(n+1) = \begin{cases} \overline{C}_j(n) + \eta(n)[C_i - \overline{C}_j(n)] & j \in A_{q(\overline{C})}(n) \\ \overline{C}_j(n) & otherwise \end{cases} \quad (4)$$

where n is the iteration number, $\eta(n)$ is the learning rate parameter which decreases monotonically with the number of iterations, $A_{q(\overline{C})}(n)$ is the neighborhood function, which decreases monotonically with the number of iterations, and $\overline{C}_j(n)$ is the best matching neuron at iteration n. The iterations stop after a predefined number of iterations or until there is no significant change in the color map $\overline{C}$.

As a result of the neighborhood function, the final quantization vector is ordered, i.e., neurons are organized into a meaningful order in which similar neurons are closer to each other in the grid than the more dissimilar ones. The ordered quantization vector is a desired property for our application since we will be using the labels of the pixels that map the colors in the image to this quantized vector.

Figure 3 shows the resulting quantized images using uniform gray-level and color quantization and non-uniform color quantization using SOFM all in pseudo-color format. The resulting quantization vector computed by SOFM is also given on the top of the third column. It can be seen that the uniform color quantization based method gives the worst result in terms of image distortion. This could be explained by the correlation between the color-channels, where this drawback is overcame by using the SOFM. The quantized image obtained using the SOFM provided a more enhanced image compared to the quantized gray level image, thus resulting more descriptive texture information.

2.3. Image Classification

For the classification, we used a combination of principal components analysis (PCA) and linear discriminant analysis followed by a Bayesian classifier. We first applied PCA and used the first k components that sum up 95% of the total variance. The class labels were determined based on the maximum likelihood method.

4. EXPERIMENTAL RESULTS

The performance of the proposed classification approach is evaluated using k-fold cross validation over a training set of 510 images. The training set is divided into five sets including equal number of samples. Each time four of these sets were used for training and the remaining one for testing. Figure 4 shows a sample scatter plot of the samples corresponding to each grade in the reduced 2D feature space with the estimated normal distributions. The separation of the grade III is clear; however there is an overlap between grades II and I. Classification results over this feature space using a Bayesian classifier is given in Table.1. The overall correct classification rate is 88.9%. We already had remarkable accuracy discriminating high-grade samples. Using the texture features, we could also differentiate the low and intermediate grades with reasonable accuracies.

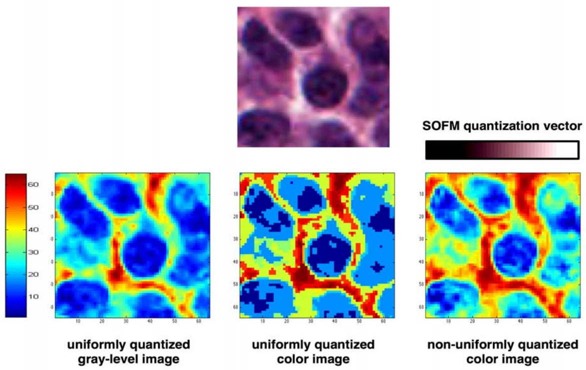

Fig.3. Quantized images using uniform gray-level and color quantization, and non-uniform color quantization using SOFM. All resulting images are shown using pseudo-coloring with the color map shown on left. The resulting quantization vector using SOFM is also given on right.

Table.1 Average classification accuracies in percentages for different grades using k-fold cross validation with k=5.

	Grade I	Grade II	Grade III
36 grade I samples	78.0	22.1	0
48 grade II samples	6.4	92.9	0.8
18 grade III samples	0	0	100

Table 2 shows how the proposed color texture approach, where we construct the co-occurrence matrix representation using SOFM based non-linear quantization improved the classification accuracies compared to that of gray level and uniform color quantization based approaches. Results were obtained using several levels of quantization leading to co-occurrence matrix sizes of 16, 32, and 64 on different methods were compared. The similar classification performances for separating high-grades are mostly due to the discrimination power of the cytoplasm to nuclei ratio feature; however we obtained considerable improvement in discriminating low- and mid-grades.

Table.2 Comparison of the classification accuracies in percentages with co-occurrence matrices using gray-level, uniform color quantization and non-linear quantization with SOFM (in boldface) from left to right, respectively. G indicates the co-occurrence matrix size.

G	Grade I	Grade II	Grade III
16	68.5/69.4/**70.2**	90.1/71.1/**91.2**	100/73.1/**100**
32	69.6/71.2/**74.0**	89.2/76.0/**91.4**	100/93.2/**100**
64	70.3/71.4/**78.0**	90.0/77.6/**92.9**	100/99.1/**100**

5. CONCLUSIONS

We introduced a novel color texture analysis method and applied it for computerized grading of follicular lymphoma images. Nonlinear quantization of color images prior to co-occurrence calculation for H&E-stained histopathological

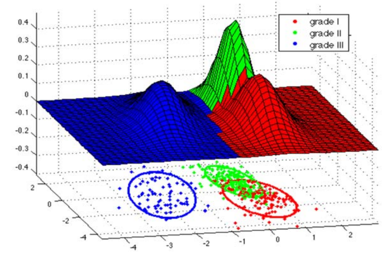

Fig.4. 2D scatter plot of the samples on the reduced feature space with the estimated probability distributions and decision boundaries determined by Bayesian classifier.

images considerably improved the classification performance, which has 100% sensitivity in identifying Grade III follicular lymphoma. The proposed color texture analysis approach is also applicable to analyze other kinds of images with limited color spectrum and this will be investigated in our future work.

6. REFERENCES

[1] Jaffe, E.S., Harris, N.L., Stein, H., and Vardiman, J.W., *Tumours of Haematopoietic and Lymphoid Tissues*, IRAC Press, Lyon, 2001.
[2] ––– "The Non-Hodgkin Lymphoma Classification Project. A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin lymphoma," *Blood*, vol. 89, pp. 3909-18, 1997.
[3] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, J. Tomasezweski, and M. Feldman, "Automated Grading of Prostate Cancer using Architectural and Textural Image Features," Proc. of *Int Symp. on Biomedical Imaging (ISBI)*, pp. 1284-1287, 2007.
[4] J. Kong, O. Sertel, H. Shimada, K.L. Boyer, J.H. Saltz, and M.N. Gurcan, "Computer-aided Grading of Neuroblastic Differentiation: Multi-resolution and Multi-classifier Approach," *Proc. of the IEEE Int.Conf on Image Processing*, pp. 525-528, San Antonio, TX, 2007.
[5] K. Masood, N. Rajpoot, H. Qureshi, and K. Rajpoot, "Hyperspectal Texture Analysis for Colon Tissue Biopsy Classification,", *Intl. Symposium on Health Informatics and Bioinformatics*, Turkey, 2006.
[6] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. on Systems, Man, and Cybernetics*, pp. 610-621, 1973.
[7] V. Arvis, C. Debain, M. Berducat, and A. Benassi, "Generalization of the Co-occurrence Matrix for Color Images: Application to Color Texture Classification," *Image Anal Stereol*, pp. 63-72, 2004.
[8] G. Paschos, "Perceptually Uniform Color Spaces for Color Texture Analysis: An Empirical Evaluation," *IEEE Trans. on Image Processing*, pp. 932-937, 2001.
[9] Kohonen, T., *Self-Organizing Maps*, Springer, Heidelberg, 1997.