Shape-Based Image Indexing and Retrieval for Diagnostic Pathology

Dorin Comaniciu [*]
Department of ECE
Rutgers University
Piscataway, NJ 08855, USA
comanici@caip.rutgers.edu

David Foran
Center for Biomedical Imaging
UMDNJ-RWJ Medical School
Piscataway, NJ 08854, USA
djf@pleiad.umdnj.edu

Peter Meer [*]
Department of ECE
Rutgers University
Piscataway, NJ 08855, USA
meer@caip.rutgers.edu

Abstract

A prototype system performing analysis, indexing and retrieval of pathology images to assist physicians in differential diagnosis of lymphoproliferative disorders is presented. Robust color segmentation is used to automatically analyse regions of interest in images of leukocytes. The shape of leukocyte nuclei, described through similarity invariant shape descriptors, represents the main attribute in the search query. Monte Carlo tests for stability and goal-directed evaluations of the system performance are also shown.

1 Introduction

Systems for the indexing, storage and retrieval of pictorial information by content have been recently developed and show promising performance [3, 6, 7]. However, databases developed for browsing photograph collections are equipped with only elementary understanding of content, most often the attributes characterizing the entire image rather than unique objects present in the image. By contrast, medical image databases demand a moderate-to-high degree of content understanding [8]. Goal-oriented indexing solutions are necessary to assist physicians in the interactive review of diagnostic images.

This paper presents the implementation of a clinical decision support system for pathology. The user delineates regions which are known to be essential for the diagnosis. Once the selections are made the system *automatically* segments the region of interest and generates attributes which describe the spatial structures within the region. These attributes in turn serve as criteria for locating, retrieving and displaying correlated clinical data.

2 Color Image Segmentation

The key component of the system is the color segmentation module. Color segmentation requires the analysis of both the *feature (color) space* and the *image domain*. Since real data typically produces color clusters of irregular geometry, only nonparametric estimation (not resorting to elliptical clusters) can provide an adequate description of the data.

In a recent paper [2] the mean shift procedure [1] has been used in the $L^*u^*v^*$ color space to detect color cluster centers for the goal of image segmentation. Mean shift is an iterative nonparametric technique for the estimation of the density gradient. The center $\mathbf{x}$ of a searching window is shifted to the average of the data points inside the window. Since the $L^*u^*v^*$ space is perceptually uniform its metric is Euclidean, and therefore it is possible to use a sphere of radius r as the searching window.

Clustering through applying the mean shift procedure to each feature point cannot be satisfactory in practical applications. First, it is very expensive, having a complexity of $O(N^2)$ where N is the number of feature points; second, the convergence over low density regions is poor, while high density regions can present plateaus without a clear local maximum. Finally, not all the clusters in the feature space have a large enough spatial support.

In the present work we extend and improve the use of the mean shift, taking into account the problems mentioned above. There are two important differences between the algorithm in [2] and the one presented here. While in [2] the clusters are removed sequentially, now they are delineated at the same time. Also, the former method used spatial information to handle nonspherical clusters, whereas in the new method the cluster geometry results from the color space analysis. The main steps of the segmentation algorithm are described below.

1. Map the image into the color space. A perceptually uniform color space is obtained by transforming the RGB vectors into $L^*u^*v^*$ vectors.

2. Define a sample set obeying distance and density constraints. To reduce the computational load, a set of points (color vectors) called the *sample set* is randomly selected from the color space. Two constraints are imposed on the points retained in the sample set. The distance between any two neighbors should not be smaller than r, the radius of the searching sphere, and the sample points should not lie in sparsely populated spheres. The latter condition is required to avoid convergence problems for the mean shift procedure. A searching sphere is sparsely populated whenever the number of points inside the sphere is below a threshold T_1. Note

[*] Supported by the NSF under the grant IRI-9530546.

that the distance and density constraints automatically determine the size K of the sample set.

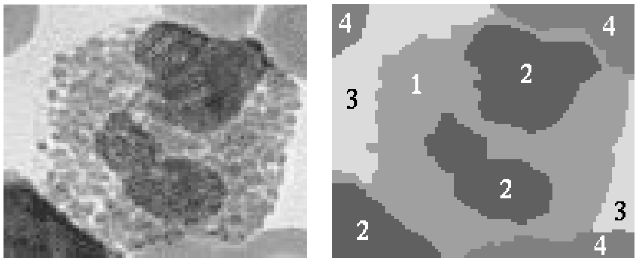
3. Apply the mean shift procedure to the sample set. A set containing K *cluster center candidates* is defined by the points of convergence of the sample points. Note the decrease in computational complexity which is now $O(KN)$ with $K \ll N$.

4. Perturb the cluster center candidates and apply to them the mean shift procedure. Since a local plateau can prematurely stop the iterations, each cluster center candidate is perturbed by a random vector of small norm and the mean shift procedure is let to converge again.

5. Derive the cluster centers from the cluster center candidates. Any subset of cluster center candidates which are sufficiently close to each other defines a *cluster center*. The cluster center is the mean of the cluster center candidates in the subset.

6. Delineate the clusters. Using a k-nearest neighbor technique over the sample set all the points not in the set are allocated to a cluster center.

7. Use spatial constraints to validate color clusters. The color vectors are mapped to the image domain. Color clusters not yielding at least one connected component greater than T_2 pixels in the image are removed and their points are reallocated. Finally, small connected components containing less than T_1 pixels are removed, and region growing is employed to allocate the unclassified pixels.



(a) (b)

Figure 1. (a) Test image.(b) Segmented image.

The segmentation parameters are the searching sphere radius r and the thresholds T_1 and T_2. It should be emphasized that all the experimental results in this paper were obtained having the segmentation parameters set to: $r = 8, T_1 = 50$, and $T_2 = 100$.

A typical leukocyte image is shown in Figure 1a. The image contains 7084 pixels and $N = 6005$ colors. The color space, the sample set ($K = 73$), the cluster center candidates, and the cluster centers are presented in Figure 2. The segmented image is shown in Figure 1b where the colors are labeled according to the extracted clusters.

The segmentation quality can also be evaluated from Figure 3 where the nucleus of each cell was delineated using the segmentation algorithm. The data set contains images of different color, sharpness, contrast, noise level and size (for convenience, they are displayed at the same size). The algorithm running time is linear with the number of pixels in the image. It takes about 2 seconds to segment a 256×256 pixel image on an Ultra SPARC 1 workstation.

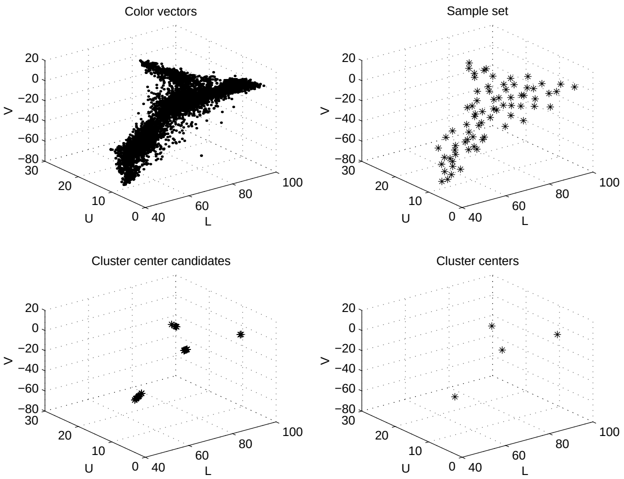


Figure 2. Various data sets involved in the segmentation. The color vectors are randomly sampled yielding the sample set which converges to cluster center candidates from which the cluster centers are extracted.

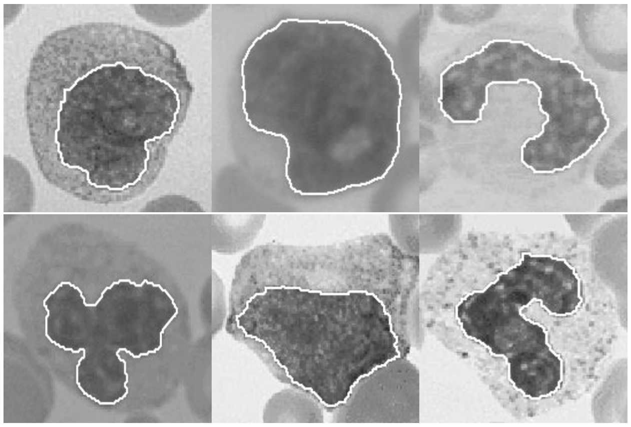


Figure 3. Nucleus segmentation (white contour) for various cell types.

3 Shape Description

Shape is an essential attribute of the nucleus and cytoplasm of the cells. The employed shape description should be *similarity invariant* (i.e., invariant to changes in location, orientation, and scale). We follow the approach in [5] to obtain the Fourier invariants, a closed contour being represented as a composition of ellipses called harmonic loci. Note that the phase information (which is important in classification tasks) is preserved by the method [5], contrary to the representations in [4, 6].

4 System Overview

The system allows the user to delineate regions of interest in the displayed images, to automatically segment the selected regions, and to formulate queries in an interactive way. The search is performed within the logical data which describe the indexed objects in terms of shape, color, scale, area, and text information.

Two pathology databases are currently used. The first one has 90 examples of the main types of leukocytes, namely band neutrophils, lymphocytes, monocytes, and polymorphonuclear leukocytes. The second database consists of 66 images which contain healthy leukocytes, and leukocytes corresponding to three particular disorders, chronic lymphocytic leukemia (CLL), follicular center cell lymphoma (FCC), and mantle cell lymphoma (MCL).

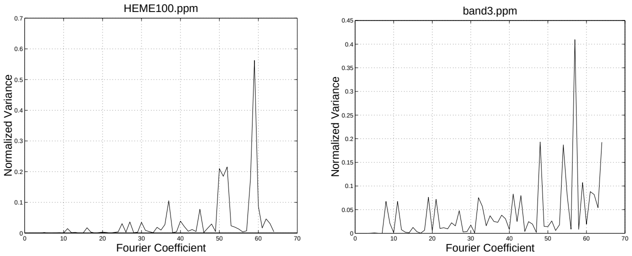


Figure 4. Normalized variance of the first 64 Fourier coefficients.

An important step in the system design was to establish the number of Fourier coefficients that can be used reliably, taking into account the uncertainty of the segmentation process. Monte Carlo experiments on five different images have been conducted to evaluate the normalized variance of each coefficient. A user has been asked to delineate 25 times a region of interest (a leukocyte) for each image, then the region was segmented, and the first 64 coefficients (16 harmonics) were determined for each nucleus. The normalized variances of the coefficients of images *HEME100.ppm* and *band3.ppm* are presented in Figure 4. The conclusion was that the segmentation is sufficiently stable for the use of the first 40 coefficients in the computation of the Euclidean distance among the nucleus shapes.

5 Retrieval Examples

Efficient image retrieval is possible by matching the shape descriptors of the query object to the descriptors within the logical data. The retrieved images are sorted in the order of their shape similarity to the query. A retrieval example presented in Figure 5. The query image is the first in the figure and the three retrieved images are band cells from our first database. It is worth to observe that the first retrieved image (the second in the figure) contains the same cell as the query image, but at a different magnification. Note that only the regions of interest are displayed and the images and cells actually have various sizes.

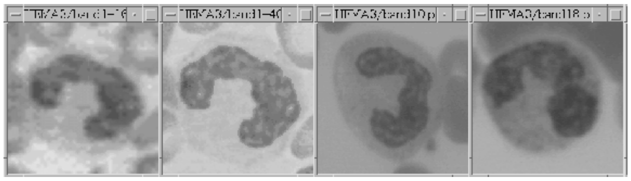


Figure 5. Shape based Band cell retrieval.

For the second database, the system uses as attributes besides the nucleus shape also the nuclear cytoplasm ratio. The first three retrieved images corresponding to the FCC query are displayed in Figure 6. A CLL query image and three retrievals are shown in Figure 7.

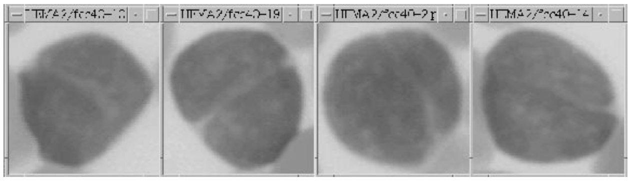


Figure 6. Nuclear cytoplasm ratio and shape based FCC cell retrieval.

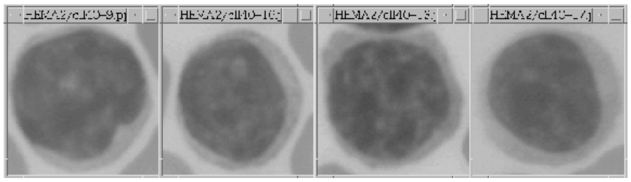


Figure 7. Nuclear cytoplasm ratio and shape based CLL cell retrieval.

References

[1] Y. Cheng. Mean Shift, Mode Seeking, and Clustering. *IEEE Trans. Pattern Anal. Machine Intell.*, 17:790–799, 1995.

[2] D. Comaniciu and P. Meer. Robust Analysis of Feature Spaces: Color Image Segmentation. *IEEE Conf. on Comp. Vis. and Pattern Recognition*, pages 750–755, Puerto Rico 1997.

[3] M. Flickner et al. Query by Image and Video Content: The QBIC System. *Computer*, 28(9):23–31, 1995.

[4] H. Kauppinen, T. Seppanen, and M. Pietikainen. An Experimental Comparison of Autoregressive and Fourier-Based Descriptors in 2D Shape Classification. *IEEE Trans. Pattern Anal. Machine Intell.*, 17:201–207, 1995.

[5] F. Kuhl and C. Giardina. Elliptic Fourier Features of a Closed Contour. *Comp. Graphics Image Process.*, 18:236–258, 1982.

[6] W. Ma and B. Manjunath. NETRA: A Toolbox for Navigating Large Image Databases. *IEEE Int'l Conf. Image Process.*, 1:568–571, Santa Barbara 1997.

[7] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-Based Manipulation of Image Databases. *Int'l. J. of Comp. Vis.*, 18:233–254, 1996.

[8] H. Tagare, C. Jaffe, and J. Duncan. Medical Image Databases: A Content-based Retrieval Approach. *J. of the American Medical Inform. Assoc*, 4:184–198, 1997.