Published in final edited form as:

IEEE Trans Inf Technol Biomed. 2009 May ; 13(3): 291–299. doi:10.1109/TITB.2008.2008801.

PathMiner: A Web-Based Tool for Computer-Assisted Diagnostics in Pathology

Lin Yang [Student Member, IEEE],

Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, 08854 and The Cancer Institute of New Jersey, New Brunswick, NJ, 08903

Oncel Tuzel [Student Member, IEEE],

Department of Computer Science, Rutgers University, Piscataway, NJ, 08854

Wenjin Chen,

The Cancer Institute of New Jersey, New Brunswick, NJ, 08903

Peter Meer [Senior Member IEEE],

Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, 08854

Gratian Salaru,

Department of Pathology and Laboratory Medicine, Robert Wood Johnson Univeristy Hospital, New Brunswick, NJ, 08903

Lauri A. Goodell, and

Department of Pathology and Laboratory Medicine, Robert Wood Johnson Univeristy Hospital, New Brunswick, NJ, 08903

David J. Foran [Member IEEE]

Center for Biomedical Imaging and Informatics, UMDNJ-Robert Wood Johnson Medical School, Piscataway, NJ, 08854 and The Cancer Institute of New Jersey, New Brunswick, NJ, 08903

Abstract

Large-scale, multi-site collaboration has become indispensable for a wide range of research and clinical activities which rely on the capacity of individuals to dynamically acquire, share and assess images and correlated data. In this paper, we report the development of a web-based system, *PathMiner*, for interactive telemedicine, intelligent archiving and automated decision support in pathology. The *PathMiner* system supports network-based submission of queries and can automatically locate and retrieve digitized pathology specimens along with correlated molecular studies of cases from "ground-truth" databases, which exhibit spectral and spatial profiles consistent with a given query image. The statistically most probable diagnosis is provided to the individual who is seeking decision support. To test the system under real-case scenarios, a pipeline infrastructure was developed and a network-based test laboratory was established at strategic sites at UMDNJ - Robert Wood Johnson Medical School, Robert Wood Johnson University Hospital, the University of Pennsylvania School of Medicine, Hospital of the University of Pennsylvania, The Cancer Institute of New Jersey, and Rutgers University. The average five class classification accuracy of the system was 93.18% based on a ten-fold cross validation on a close dataset containing 3691 imaged specimens. We also conducted prospective performance studies with the *PathMiner* system in real applications in which the specimens exhibited large variations in staining characters compared with the training data. The average five-class classification accuracy in this open-set experiment was 87.22%. We also provide the comparative results with the previous literature and the *PathMiner* system shows superior performance.

Index Terms

Segmentation; classification; content based image retrieval; computer aided diagnostics

I. Introduction

While blood cells are often differentiated based upon traditional morphological characteristics, the subtle visible differences exhibited by some lymphomas and leukemias gives rise to a significant number of false negatives during routine screening by medical technologists. In many cases the differential diagnosis can only be rendered after immunophenotyping, and molecular or cytogenetic study of the cells involved. The additional studies are expensive, time consuming, and usually require fresh tissue which may not be readily available. In addition it occurs too late in the diagnostic pathway to impact significantly on the frequency of false negatives. While it would be impractical to immunophenotype every sample that is flagged by complete blood count (CBC), passing the specimen through a reliable, image-based screening system could potentially reduce cost and patient morbidity.

Developing strategies which transform complex diagnostic reasoning into reliable algorithmic procedures remains a very active field of research [1], [2], [3] with several projects focusing on clinical and anatomic pathology. These include the Pathex framework and the Pathex/Red system [4] which was developed at Ohio State University to assist pathologists in evaluating laboratory data, the ECLIPS [5] system developed at the University of Illinois Urbana, and the PathFinder project which was designed and developed at the University of Southern California and Stanford to provide assistance in rendering diagnostic decisions in anatomic pathology [6]. The PathFinder system provides a differential diagnosis based upon the initial histological features that are observed by the pathologists, and provides suggestions as to what additional histological features are most likely to narrow the differential diagnosis, thus helping to screen for incompatible observations for specific diseases.

Technologies that can adequately capture the visual essence of 2-D and 3-D objects rely on strategies and principles which have grown out of research in image analysis, pattern recognition, and database theory. Several general purpose content-based image retrieval (CBIR) systems have been developed which exploit these technologies such as the IBM QBIC system [7], the Photobook system [8], the WBIIS system [9], the Blobworld system [10] and the IRM system [11]. Recently there has been increased interest and efforts applied to utilizing CBIR in medical applications [12]. For example, the Pittsburgh Supercomputing Center developed a system which utilizes global characteristics of images to provide a measure of gleason grade of prostate tumors [13]. The wavelet technology and Integrated Region Matching (IRM) distances are used in [9] for characterizing pathology images.

In this paper, we describe the *PathMiner* system, which can automatically scan, index, segment and classify blood smear specimens. Five different classes of blood cells were used to evaluate the system performance. These include Benign, chronic lymphocytic leukemia (CLL), mantle cell lymphoma (MCL), follicular center cell lymphoma (FCC) and acute leukemia (ALL and AML).

The remainder of the paper is organized as follows: Section 2 introduces the system architecture of *PathMiner* and provides detailed description of all its components. Segmentation, classification and image rank retrieval algorithms are presented in Section 3. Section 4 provides the experimental results and Section 5 concludes the paper.

II. System Description

PathMiner is a web-based system for interactive telemicroscopy automated analysis and interpretation of digitized pathology and correlated data. The major components of *PathMiner* are a distributed telemicroscopy subsystem (DT), an intelligent archival (IA) subsystem and an image guided decision support (IGDS) subsystem.

The distributed telemicroscopy (DT) subsystem of *PathMiner* enables primary users to communicate remotely and share information. The DT subsystem features auto-focusing, shared graphics and text messaging. It is a crucial component for physicians to exchange diagnostic opinions. The intelligent archiving (IA) subsystem performs automatic and remote control of the microscope. The image guided decision support (IGDS) subsystem enables individuals to submit query images originating from local and remote computers. The IGDS automatically segments and classifies the query image. It also provides content based rank retrieval results after probing a "ground-truth" database.

In addition to the three major components, DT, IA and IGDS, a global control middleware (GCM) was developed to coordinate activities among each of the components. In Figure 1, we show an overview of the system architecture. Because each subsystem can be launched independently, multiple components can be pipelined with one another to accommodate complex image analysis tasks.

Figure 2 shows the system architecture when the IA subsystem, IGDS subsystem and GCM module are pipelined to perform unsupervised specimen analysis. Subsystems can be run on single machine or distributed nodes using TCP/IP as the communication protocol. In Figure 2, the solid lines represent the data flow throughout the course of the entire operation and the dotted lines represent the communication paths between the system and human operators. The operator is only required to place a specimen slide on the microscope stage (the blue dotted arrow pointing to the IA system).

A. Intelligent Archiving System (IA)

Prior to the development of the Intelligent Archiving (IA) system, it took a large amount of time for pathologists to index new cases into the "ground-truth" database. Specimens were first reviewed under the microscope at low magnification while cells of interest were interactively brought into focus at high magnification. When a cell of interest came into the view the pathologist invoked software to signal the digital camera to digitize the microscopic field. In order to index the imaged cells, each cell was individually loaded into the IGDS system and then the resulting image and features were populated into an Oracle 10g database.

The IA subsystem was designed for automatic imaging and indexing procedures. This was implemented using a computer assisted microscopy (CAM) server module which reliably translates image-based coordinates into physical coordinates along the optical paths of the robotic microscope's objective lenses. The CAM module features intelligent control of the microscope which enables it to coordinate activities among the primary devices. The CAM module also enables the imaging system to precisely estimate the region of interest (ROI) of a microscopic object, e.g. a cell.

The intelligent control provided by the computer assisted microscopy module allows the IA system to perform unsupervised detection, imaging and indexing of candidate lymphocytes into "ground-truth" databases. The work flow and the IA user interface are illustrated in Figure 3. The CAM module first directs the robotic scope to perform an unsupervised pilot scan over the specimen at low resolution (10 $\times$), which captures slightly overlapping frames

and stitches them together generating a whole image map. Color filtering of the image map is subsequently performed in *Luv* color space to detect leukocytes while spatial constraints were applied to eliminate artifacts. Exact stage coordinates of each candidate cells are extracted and used to direct the robotic scope to systematically image the leukocytes at high resolution while simultaneously performing segmentation and image feature extraction.

Since the captured cells may include errors, a second level of filtering is implemented to reject candidate cells whose feature profiles are inconsistent with that of a lymphocyte. The rejection filter (*LymphGate*) is based on cell area and the roundness factor, which is computed as

$$roundness = \frac{1}{eccentricity} = \frac{perimeter^2}{4\pi * area} \quad (1)$$

Those cells that fall outside of empirically derived limits are rejected. The remaining imaged lymphocytes and their corresponding image-based feature metrics are sent to the IGDS system for further processing.

The IA subsystem has a client-server design. The client side can be launched from the IGDS subsystem which was developed using an Olympus AX70 microscope equipped with a prior 6-way robotic stage and motorized turret. The minimum requirements for server workstations consist of a standard Pentium IV computer, equipped with 512 M bytes of RAM.

B. Image Guided Decision Support System (IGDS) and Global Control Manager (GCM)

The overall system architecture of the IGDS subsystem is shown in Figure 4. The client Graphic User Interface (GUI) of the IGDS allows two types of queries. In the single mode, the client GUI of the IGDS allows users to load the input query image into the IGDS system. In the hybrid mode depicted in Figure 2, a pre-selected ROI image provided by the IA system is automatically fed to IGDS for segmentation and classification. In the client processing mode, both the nucleus and the cytoplasm of each cell in the ROI are segmented using the newly developed robust color GVF (RCGVF) algorithm [14].

The IGDS supports two different sets of morphometric features which can be extracted from each cell's nucleus and cytoplasm once the image segmentation step has been completed. The first set consists of feature measurements for shape as described by elliptical Fourier descriptors; color as defined in *Luv* color space; nuclear and cytoplasmic area; the cytoplasm/nuclear ratio and the texture measurements. The image metrics which are generated are then automatically inserted into the "ground-truth" database and made available for queries. When individuals confront a difficult or ambiguous case they can query the updated "ground-truth" database for decision support. A weighted combination of the features generated for the unclassified cell or cells are automatically compared to those within the "ground-truth" database. A *k*-nearest Neighbor classifier is used to determine the degree of similarity with the query image.

The IGDS subsystem also features the option to utilize textons [15], which have been shown to be effective for applications where chromatin patterns and granularity of the cell contain discriminative information. Textons are defined as conspicuous repetitive local features that human perceive as being discriminative among textures. The computational model for textons, introduced by Leung and Malik [16], are cluster centers in a feature space which are generated in responses to a fixed set of filter banks. The IGDS subsystem extracts texton measurements for the nucleus and cytoplasm of each cell. The algorithms used to perform

unsupervised segmentation, content based image retrieval and classification are described in Section 3.

As the query is generated, it is sent to the IGDS system through a global control manager (GCM). The feature measurements of the unknown samples are compared with the signatures which have been stored in the “ground-truth” database. Several B+ trees are generated to increase the speed of accessing the data in the Oracle database. By default, the digitized specimens of the first eight ranked retrievals, the statistically most likely diagnosis based on the classification and all correlated clinical information (e.g. molecular and protein profiles) are returned to the individual who is seeking decision support.

The global control manager (GCM) module is generally used by the system administrator of the *PathMiner* system. It has three main functions: 1) a graphic representation of all log-ins and a list of each individual’s operations; 2) a graphical representation which indicates the use of the “ground-truth” databases, such as the number of cases stored, the number of imaged specimens, the location of distributed databases and a measure of the disk usage; 3) an administration tool which can be used to control the privileges of users. Figure 5 shows the graphic user interface of IGDS and GCM.

If the user has the proper authority to populate the ground-truth database, e.g. certified pathologists, they can launch the unsupervised indexing process. For quality control, all the segmentation and analysis results are made available for reviewing by certified pathologists prior to their becoming integrated into the core “ground-truth” database. Both the IGDS and GCM were developed using Java 1.5.0 for the purpose of platform independence. Oracle 10g serves as the database engine.

C. Distributed Telemicroscopy System (DT)

The Distributed Telemicroscopy (DT) subsystem enables individuals located at disparate clinical and research sites to engage in interactive consultation. It allows primary users to control the specimen stage, objective lens, light levels and focus of robotic microscopes, remotely. A digital representation of the specimen is continuously broadcast to all session participants. Primary user status can be passed among session participants as a software token. The system features shared graphical pointers, text-messaging capability, and automated database management. It is a crucial component for participants to exchange the diagnostic opinions while they are using the Image Guided Decision Support subsystem. A snapshot of the Distributed Telemicroscopy (DT) client interface is shown in Figure 6.

III. Image Analysis Algorithms

The *PathMiner* system contains the algorithms for automatic segmentation, content based image retrieval and classification. The algorithms used in the system are described in the following section.

A. Image Segmentation

In order to perform the content based image retrieval and classification of the imaged cells, the first crucial step is segmentation. For the specimens under study, both the nuclei and the cytoplasm of imaged cells exhibit distinguishable information which is important for classification. A robust color GVF (RCGVF) snake [14] was developed specifically to segment the nuclei and the cytoplasm of the cells. We first apply a L_2E robust estimation [17] to provide a rough estimation of the outer boundaries of the cells inside the region of interest (ROI). A gradient vector flow (GVF) snake [18] using Luv [19, Sec. 8.4] color gradients is further applied to extract the objects from the background. The proposed method

can segment a 255×255 image within one second on a Pentium PC with 1.5G Hz CPU and 512M memory.

Figure 7a shows good, fair and poorly representative imaged cells from each class which were stained with hematoxylin and eosin. In our testing set which contains 3691 cell images, about 35% belongs to good images, 40% is the fair images and the rest 25% belongs to the poorly representative images. The algorithm is able to provide satisfactory performance even when confronted with images exhibiting weak contrast and subtle edges. We obtained an average accuracy 90.1% on the entire database. Some segmentation results are shown in Figure 8. For more details about the color gradient and robust color gradient vector flow (GVF) snake we refer readers to [14].

The segmentation algorithm is further extended to manage touching cells segmentation [20] and tested on a dataset which contains 207 touching cells (shown in Figure 7b). In the touching cell dataset we obtained an average segmentation accuracy 88.9%. Since the watershed algorithm [21] is widely accepted for touching object segmentation and has been successfully used in segmenting histopathology images [22]. We compared our touching cell segmentation method with watershed using the 207 touching cell image dataset and shown the results in Table I. The 80% column in Table I represents the sorted 80% highest accuracy of all the results, and is commonly used by pathologists to assess the segmentation accuracy. The experimental results demonstrate the superior performance of the new segmentation algorithm.

B. Content Based Image Retrieval

The shape of the nuclei, the texture and area of both nuclei and cytoplasm, and the ratio between cytoplasm and nuclei area are used as the features to measure the similarities of image rank retrieval. The shape of the cytoplasm is not utilized since it is often distorted by its neighboring cells.

Instead of using a complex shape model, the Elliptic Fourier Descriptor (EFD), which was shown to be successful in [23], is chosen to model the shape of the nuclei. There are several advantages of choosing EFD: 1) the EFD has a simple histogram-like representation. In our system we use the first 32 ($4 * 8$) coefficients; 2) the normalized EFD is invariant to rotation, translation and scaling; 3) the close contour reconstructed from EFD is always closed.

EFD is the Fourier expansion of the chain coding. Assume we have M points on the close contour. Following the approach of Kuhl and Giardina [24], the EFD coefficients of the n th harmonic are:

$$\begin{aligned} a_n &= \frac{S}{2n^2\pi^2} \sum_{i=1}^M \frac{\Delta x_i}{\Delta s_i} \left[\cos \frac{2n\pi s_i}{S} - \cos \frac{2n\pi s_{i-1}}{S} \right] \\ b_n &= \frac{S}{2n^2\pi^2} \sum_{i=1}^M \frac{\Delta x_i}{\Delta s_i} \left[\sin \frac{2n\pi s_i}{S} - \sin \frac{2n\pi s_{i-1}}{S} \right] \\ c_n &= \frac{S}{2n^2\pi^2} \sum_{i=1}^M \frac{\Delta y_i}{\Delta s_i} \left[\cos \frac{2n\pi s_i}{S} - \cos \frac{2n\pi s_{i-1}}{S} \right] \\ d_n &= \frac{S}{2n^2\pi^2} \sum_{i=1}^M \frac{\Delta y_i}{\Delta s_i} \left[\sin \frac{2n\pi s_i}{S} - \sin \frac{2n\pi s_{i-1}}{S} \right] \end{aligned} \quad (2)$$

where $s_i = \sum_{j=1}^{i(M)} \Delta s_j$, $\Delta s_i = \sqrt{(\Delta x_i)^2 + (\Delta y_i)^2}$, $\Delta x_i = (x_i - x_{i-1})$, $\Delta y_i = (y_i - y_{i-1})$. The Δx_i and Δy_i are the changes in the x and y projection of the chain code at the i th contour point.

In addition to shape, texture is also used for content based image retrieval. The calculation of texture feature vector (the texton histogram) will be described in detail in the next section. The final similarity metric was defined as the weighted distance of all the features including shape, texture, area and color

$$D = \sum_{i=1}^n w_i f_i \quad (3)$$

where n is the number of features, w_i is the corresponding weights of each feature and f_i is the i -th Euclidean distance between the feature vector of the candidate and the "ground-truth" targets in the database.

C. Imaged Cell Classification

In this section, we describe the methods used to classify the digitized imaged blood cells. Texton histograms were used as feature measures to classify the staining profiles of the nuclei and cytoplasm of the imaged blood cells. Because the feature vectors lie in a relatively high dimensional space, the maximal margin classifier, support vector machine (SVM), is used to classify the cell images.

1) Texton Histogram—Texture can be characterized through clusters which are organized patterns of the basic elements. Current state-of-the-art texture research is based on characterizing textures using responses to sets of linear filters. This approach has been successfully used in several fields of research including classification, segmentation and synthesis [25], [26], [16], [27].

We utilize texture based features to represent each imaged cell. Following segmentation, the images were converted to gray scale and normalized such that the mean was zero and standard deviation was one. Since the images were acquired under different experimental conditions, normalization was applied to minimize the variability.

Pixels inside each cell cytoplasm and nuclei were convolved with the M8 filter bank [27] consisting of 38 filters (shown in Figure 9). The filters used in this filter bank were a Gaussian and a Laplacian of Gaussian both with $\sigma = 10$ pixels (these filters have rotational symmetry), an edge filter at 3 scales (scale values) = (1,3), (2,6), (4,12) and a bar filter at the same 3 scales. Among the oriented edge and bar filters only the maximum filter response is retained at each scale. As a result each pixel in the image was represented as an eight dimensional feature vector.

The texture of the cytoplasm and nuclei were analyzed independently. A few random images were selected from each class. The filter responses were clustered using k -means clustering algorithm ($k = 45$ which is learned in an offline process using a training set and held constant throughout the experiments). The clustering is performed separately for pixels inside the nuclei and cytoplasm. Since the size and the variability of the cytoplasm texture is less than that for the nuclei, half the number of clusters were generated for the cytoplasm as compared to the nuclei (30 and 15 cluster centers for nuclei and cytoplasm texture respectively). The cluster centers, called textons, were used to generate a texton library. The appearance of each blood cell image was modeled by a compact quantized description called texton histograms. Texton histograms are created by assigning each pixel filter response in the image to its closest texton in the texton library that was generated. This was calculated using

$$h(i) = \sum_{j \in I} \text{count}(T(j)=i) \quad (4)$$

where I denotes the cell image, i is the i th element of the texton dictionary, $T(j)$ returns the texton assigned to pixel j . In this way, each cell image was modeled as a texture modes distribution, the texton histogram. Each image was mapped to one point in the high dimension space R^d , where $d = K = 45$ is the number of textons.

Given an arbitrary testing imaged cell, the pixels inside the cytoplasm and nuclei were filtered and the responses were quantized to the nearest texton. Using the learned texton libraries, each cell was represented by its texton histogram. The size of the cytoplasm and nuclei are important for the analysis, therefore we did *not* normalize the histogram and each bin of the histogram was equal to the number of occurrence of the texton in the image.

2) Classification—The *support vector machine* (SVM) was first introduced in [28] for binary classification problem. The strategy is to construct the linear decision boundaries in a large transformed version of the original feature space. The SVM simultaneously minimizes the empirical classification error and maximizes the geometric margins by minimizing the regularization penalty

$$\frac{1}{2} \|\mathbf{w}\|^2, \text{ subject to } y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 \geq 0 \quad (5)$$

When the examples are not linearly separable, the optimization can be modified by adding a penalty for violating the classification constraints. This is called *soft margin SVM* which minimizes

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i, \text{ subject to } y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 + \xi_i \geq 0 \quad (6)$$

where ξ_i are called slack variables which store the deviation from the margin and C is the soft penalty to balance the training errors and margins. In (5) and (6), $\mathbf{w}$ is the slope of the decision hyperplane and w_0 is the offset. The $\mathbf{x}_i$ denotes the feature vector, and y_i is the ground true labels. We minimize (6) by maximizing the dual problem of (6) which involve a feature mapping $\phi(\mathbf{x})$ through an inner product. The inner product can be evaluated without ever explicitly constructing the feature vectors $\phi(\mathbf{x})$ but through a kernel function $\kappa(\mathbf{x}, \mathbf{x}')$. In *Pathminer*, we proposed to use a linear kernel defined as

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}' \quad (7)$$

where $\mathbf{x}$ represents the feature vector, which is the the texton histogram in our case.

In order to extend the method to multi-class problem, we constructed a binary classifier for each combination of the classes (one-against-one SVM) [29]. The label of a test example was predicted by the majority voting among the classifiers. A more detailed discussion of our SVM based classification can be found in [30].

IV. Experiments

The specimens were prepared using standard protocols where a drop of blood was placed on the glass slide and smeared into a thin film using an automatic slide maker device. The smear was then air-dried and stained using the standard staining protocols for preparing hematology specimens. According to the number of acid and basic groups present, cell components take up the dyes from the mixture in a variety of proportions. Different cells exhibited different hues depending on their composition (in proteins, amino acids, enzymes, etc.). However, for a particular cell type the staining quality was generally stable.

The test platform for the experiments consisted of an Intel-based workstation interfaced with a high-resolution Olympus DP70 camera equipped with 12-bit color depth on each color channel and 1.45 million pixel effective resolution. The system also includes a single 2/3 inch CCD digital camera, an Olympus AX70 microscope equipped with a Prior 6-way robotic stage, motorized objective turret and a magnification changer.

Traditional classification accuracy reported in the literature is usually based on a single dataset, which we refer to as a closed dataset. The accuracy is then tested using cross validation or by taking a portion of the dataset as the testset. In our experiments, we evaluated the performance using two independent datasets. The first dataset (close dataset) was used for training and a ten-fold cross validation result was reported. The second dataset, referred to as the open dataset, contained 1200 imaged cells exhibiting large variations in staining characteristics which were used only for testing.

The cell types in both two datasets included a mixed set of mantle cell lymphoma (MCL), chronic lymphocytic leukemia (CLL), follicular center cell lymphoma (FCC), acute leukemia and benign. The imaged cells were collected from the Hospital of the University of Pennsylvania, Philadelphia, PA, Robert Wood Johnson University Hospital, New Brunswick, NJ, and City of Hope National Medical Center, Duarte, CA.

The closed dataset contained 86 hematopathology cases: 18 MCL, 20 CLL, 9 FCC, 39 acute leukemias and 19 benign cases. For each case 10–90 cell images were generated. In total there were 3691 images taken from 105 different cases. The data distribution and the final classification accuracy was reported based on a ten fold cross validation. The result is listed in Figure 10. Note that one of the largest errors is due to the ambiguity between MCL and CLL, which is consistent with the performance of pathologists. The lower classification accuracy of FCC is due to the fact that there were less FCC training samples. The average five class classification accuracy was 93.18% based on ten fold cross validation in this closed dataset.

We also compare our approach with the method of [23]. The problem considered in their experiments contained only four classes (Normal, MCL, FCC, CLL) using only 261 specimens. The testing was performed by adopting tenfold cross validations. The confusion matrix is shown in Table II. In order to achieve fair comparison, we also performed ten-fold cross validations and presented the cell classification results in Table III. It is obvious that even though we are solving a more difficult problem (with one more class of disease), our system still achieved significantly better result than [23] except for the FCC class. We note that the reason for this is because there were only 20 FCC cells in [23].

The open dataset contained 30 new cases taken from new specimens. It contained 5 MCL, 6 CLL, 3 FCC, 8 acute leukemia and 8 Benign cases. In each case 40 images were generated. In total there were 1200 images digitized from these 30 mixed sets of cases. None of the images were ever shown to the system until testing and there existed obvious variations in the staining characteristics of specimens across the institutions, which were introduced from

differences in manufacturing of the dyes, choices in automated stainers and the overall intensity variations. The data distribution and the classification results are shown in Figure 11. The average five class classification accuracy was 87.22%. The lower accuracy compared with the results on the closed dataset is due to the new inter-class similarities and intra-class variations which were never seen during the training.

In Figure 12 we show some representative classification samples. The left four columns are the correct classification samples using *PathMiner* and the right fifth column shows the failed samples. The first row is the Benign cell class, where the last one is misclassified as CLL. The second row represents CLL where in the last cell image is misclassified as MCL. The third row is MCL and the last cell image is misclassified as Benign. The fourth row is FCC in which case the last image is misclassified as MCL. The last row is acute leukemia and the last image is misclassified as FCC. In Figure 12 we can see that there exists inter-class similarities and intra-class variations which make the multi-class cell classification a quite challenging problem.

V. Conclusion

Examination of peripheral blood smears represents a major activity for a hematology laboratory. When suspicious slides are flagged for review by a technician a lengthy process ensues, finalized either by reporting the lymphocytes as normal/reactive or by recommending immunophenotyping by flow cytometry. Computer assisted diagnostics (CAD) can reduce the work-load of technicians and pathologists. In this paper, we have developed a complete CAD system for performing computer-assisted assessment of imaged pathology specimens and have conducted a large-scale set of experiments (135 cases with 4891 imaged cells in total) using both closed-set and open-set performance analysis. We also introduced a newly developed pipeline infrastructure for the system and demonstrated its usage in unsupervised specimen analysis. Over the past few years our experiments have progressed from single-cell and multi-cell analysis to the evaluation of complex histological sections. *PathMiner* was developed with modular design with a flexible workflow so that with a minimal amount of modifications it can be utilized to support additional applications. We have already begun to use the *PathMiner* framework as the core platform for developing a high-throughput microscopy system for performing comparative analysis of expression patterns in immunostained tissue microarrays[15].

Acknowledgments

This research was supported, in part, by a grant from The Cancer Institute of New Jersey and NIH contracts 5R01LM007455-03 and 5R01EB003587-03 from the National Library of Medicine and the National Institute of Biomedical Imaging and Bioengineering, respectively. UMDNJ also wants to thank and acknowledge IBM for providing free computational power and technical support for this research through World Community Grid.

References

1. Debeir O, Decaestecker C, Pasteels JL, Salmon I, Kiss R, Van HP. Computer-assisted analysis of epiluminescence microscopy images of pigmented skin lesions. *Cytometry*. 1999; 37(4):255–266. [PubMed: 10547610]
2. Cenci M, Nagar C, Vecchione A. PAPNET-assisted primary screening of conventional cervical smears. *Anticancer Research*. 2000; 20(5):3887–3889. [PubMed: 11268471]
3. Kok MR, van Der Schouw YT, Boon ME, Grobbee DE, Kok LP, Schreiner-Kok PG, van der Graaf Y, Doornwaard H, van den Tweel JG. Neural network-based screening (nns) in cervical cytology: No need for the light microscope? *Diagnostic Cytopathology*. 2001; 24(6):426–434. [PubMed: 11391826]

4. Smith J, Svirbely J, Evans C. RED: A red-cell antibody identification system. *Journal of Medical System*. 1985; 9:121–137.
5. Thursh DR, Marby F, Levy A. Computers and video discs in pathology education: ECLIPS as an example of one approach. *Human Pathology*. 1986; 17(1):216–218. [PubMed: 3949342]
6. Nathwani BN, Clarke K, Lincoln T, Berard C, Taylor C. Computers and video discs in pathology education: ECLIPS as an example of one approach. *Human Pathology*. 1997; 28(9):117–121. [PubMed: 9023389]
7. Faloutsos C, Equitz W, Flickner M, Niblack W, Petkovic D, Barber R. Efficient and effective querying by image content. *Journal of Intelligent Information System, Integrated Artificial Intelligence and Database Technologies*. 1994; 3(3):231–262.
8. Pentland A, Picard RW, Scarloff S. Photobook: Content based manipulation of image databases. *International Journal of Computer Vision*. 1996; 18:233–254.
9. Wang JZ, Wiederhold G, Firschein O, Sha XW. Content-based image indexing and searching using daubechies wavelets. *International Journal of Digital Libraries*. 1998; 1(4):311–328.
10. Carson C, Belongie S, Greenspan H, Malik J. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans Pattern Analysis and Machine Intelligence*. 2002; 25(8):1027–1037.
11. Li J, Wang JZ, Widerhold G. Irm: Integrated region matching for image retrieval. *Proc of the ACM Multimedia*. 2000; 1(4):147–156.
12. Schnnorrenberg F, Pattichis CS, Schizas CN, Kyriacou K. Content-based retrieval of breast cancer biopsy slides. *Technol Health Care*. 2000; 8(5):291–297. [PubMed: 11204175]
13. Wetzel AW. Computational aspects of pathology image classification and retrieval. *Journal of Supercomputing*. 1997; 11(1):279–293.
14. Yang L, Meer P, Foran D. Unsupervised segmentation based on robust estimation and color active contour models. *IEEE Trans Information Technology in Biomedicine*. 2005; 9:475–486.
15. Yang, L.; Chen, W.; Meer, P.; Salaru, G.; Feldman, MD.; Foran, DJ. High throughput analysis of breast cancer specimens on the grid. *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention*; 2007.
16. Leung, T.; Malik, J. Recognizing surfaces using three-dimensional textons. *Proc. IEEE International Conference on Computer Vision*; 1999. p. 1010–1017.
17. Scott DW. Parametric statistical modeling by minimum integrated square error. *Technometrics*. 2001; 43:274–285.
18. Xu C, Prince JL. Snakes, shapes and gradient vector flow. *IEEE Trans Image Processing*. 1998; 7(3):359–369.
19. Wyszecki, G.; Stiles, WS. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. 2. Wiley; 1982.
20. Yang, L.; Tuzel, O.; Meer, P.; Foran, DJ. Touching cells segmentation in hematologic specimens using concave vertex graph. *Proc. International Conference on Medical Image Computing and Computer Assisted Intervention*; 2008.
21. Moga AN, Gabbouj M. Parallel marker-based image segmentation with watershed transformation. *Journal of Parallel and Distributed Computing*. 1998; 51(1):27–45.
22. Adiga PSU, Chaudhuri BB. An efficient method based on watershed and rule-based merging for segmentation of 3D histopathological images. *Pattern Recognition*. 2001; 34(7):1449–1458.
23. Comaniciu D, Meer P, Foran D. Image-guided decision support system for pathology. *Machine Vision and Applications*. 1999; 11:213–224.
24. Kuhl FP, Giardina CR. Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing*. 1982; 18:236–258.
25. Cula O, Dana K. 3D texture recognition using bidirectional feature histograms. *International Journal of Computer Vision*. 2004; 59(1)
26. Heeger, D.; Bergen, J. Pyramid-based texture analysis/synthesis. *Proc. ACM International Conference and Exhibition on Computer Graphics and Interactive Techniques*; 1995. p. 229–238.
27. Varma M, Zisserman A. Classifying images of materials: Achieving viewpoint and illumination independence. *Proc European Conference on Computer Vision*. 2002; 3:255–271.

28. Cortes C, Vapnik V. Support vector networks. *Machine Learning*. 1995; 20:1–25.
29. Hsu C, Lin C. A comparison of methods for multiclass support vector machines. *IEEE Trans on Neural Networks*. 2002; 13:415–425.
30. Tuzel O, Yang L, Meer P, Foran DJ. Classification of hematologic malignancies using texton signatures. *Pattern Analysis and Applications*. 2007; 10:277–290. [PubMed: 19890460]

Fig. 1.
The system architecture of the *PathMiner* system.

Fig. 2.

The pipeline structure of the automatic decision support procedure for classifying the digitized blood cells. The Intelligent Archiving (IA) system is used to control a microscope to automatically capture the cell images. The Image Guided Decision Support (IGDS) is the core component for performing image analysis and content based image retrieval. The Global Control Middleware (GCM) is designed to enable the system administrator to adjust the workload and the authority of the users. The Distributed Telemicroscopy (DT) is used for the communication among users.

Fig. 3.

The work flow (left panel) and the user interface (right panel) of the intelligent archiving (IA) system. The pre-scanned result of an image patch under 10× magnification is on the left-top sub plot of the right panel.

Fig. 4.
The system architecture of image guided decision support (IGDS) subsystem in *PathMiner*.

Fig. 5. The graphic user interface of image guided decision support (IGDS) system is shown on the left and the global control manager (GCM) is shown on the right. The dotted line in the GCM represents the requirement issued from the user, the solid line denotes the requirement is fulfilled by GCM.

Fig. 6.
The graphic user interface of distributed telemicroscopy (DT) system.

Fig. 7.
 (a) The representative cell images. From left to right: chronic lymphocytic leukemia (CLL), mantle cell lymphoma (MCL), follicular center cell lymphoma (FCC) and acute leukemia (ALL and AML, respectively). From top to down: good, fair and poorly representative cells.
 (b) Some samples of touching cells

Fig. 8.
The segmentation results on weak boundaries cases and touching cells cases

Fig. 9.
The M8 filter bank.

Fig. 10.

The data distribution and the classification results using one-against-one SVM and tenfold validation on the first dataset, which contains 3691 testing images.

Fig. 11.

The data distribution and the classification results on the open dataset which contains 1200 testing images. The SVM classifiers are trained using the closed dataset and never retrained for open dataset test.

Fig. 12.

The multi-class classification results using one-against-one SVM. The left four columns are correct classified samples, and the right fifth column shows the failed images.

TABLE I
The segmentation accuracy (%) using the watershed algorithm and our proposed method.

	Mean	Variance	Median	Min	Max	80%
Watershed	74.3	9.8	75.1	65.4	82.7	72.9
RCGVF	88.9	5.1	90.2	75.2	95.5	87.1

TABLE II

The confusion matrix (ten-fold cross validation) using the algorithm proposed in [23]

	Normal	CLL	MCL	FCC	No Decision
Normal	73.0	13.4	0	12.0	1.6
CLL	7.0	83.9	7.1	2.0	0
MCL	0	13.6	83.3	1.4	1.7
FCC	5.0	2.5	0	90.0	2.5

TABLE III

The confusion matrix (ten-fold cross validation) using the one-against-one Support Vector Machine

	Normal	CLL	MCL	FCC	Acute
Normal	96.2	3.4	0.4	0	0
CLL	2.9	90.4	3.9	2.8	0
MCL	1.5	6.0	83.6	1.5	7.4
FCC	1.9	9.7	6.2	81.4	0.8
Acute	0	0	1.1	0	98.9