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

Lin Yang, *Student Member, IEEE*, Oncel Tuzel, *Member, IEEE*, Wenjin Chen, Peter Meer, *Senior Member, IEEE*, Gratian Salaru, Lauri A. Goodell, and David J. Foran, *Member, IEEE*

***Abstract*—Large-scale, multisite collaboration has become indispensable for a wide range of research and clinical activities that 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 that 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 the University of Medicine and Dentistry of New Jersey—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 tenfold 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*—Classification, computer-aided diagnostics, content-based image retrieval, segmentation.**

I. Introduction

While blood cells are often differentiated based on traditional morphological characteristics, the subtle visible differences exhibited by some lymphomas and leukemias give 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 that 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, passing the specimen through a reliable, image-based screening system could potentially reduce cost and patient morbidity.

Developing strategies that transform complex diagnostic reasoning into reliable algorithmic procedures remains a very active field of research [1]–[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 on 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 that 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 that 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

Manuscript received January 23, 2008; revised June 1, 2008 and October 14, 2008. First published January 20, 2009; current version published May 6, 2009. This work was supported in part by a grant from The Cancer Institute of New Jersey, by the National Library of Medicine under Contract 5R01LM007455-02, by the National Institute of Biomedical Imaging and Bioengineering under Contract 5R01EB003587-03.

L. Yang is with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854 USA, and also with The Cancer Institute of New Jersey, New Brunswick, NJ 08903 USA (e-mail: linyang@eden.rutgers.edu).

O. Tuzel was with the Department of Computer Science, Rutgers University, Piscataway, NJ 08854 USA. He is now with Mitsubishi Electric Research Laboratories, Cambridge, MA 02139 USA.

W. Chen is with the Center for Biomedical Imaging and Informatics, University of Medicine and Dentistry of New Jersey, Piscataway, NJ 08854 USA.

P. Meer is with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854 USA.

G. Salaru is with the Department of Pathology and Laboratory Medicine, Robert Wood JohnsonMedical School/University of Medicine and Dentistry of New Jersey, New Brunswick, NJ 08903 USA.

L. A. Goodell is with the Department of Pathology and Laboratory Medicine, Robert Wood JohnsonMedical School/University of Medicine and Dentistry of New Jersey, New Brunswick, NJ 08903 USA, and also with the Immunohistochemistry Core Research Laboratory, The Cancer Institute of New Jersey, New Brunswick, NJ 08903 USA.

D. J. Foran is with the Center for Biomedical Imaging and Informatics, University of Medicine and Dentistry of New Jersey—Robert Wood Johnson Medical School, Piscataway, NJ 08854 USA, and also with The Cancer Institute of New Jersey, New Brunswick, NJ 08903 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2008.2008801

interest and efforts applied toward utilizing CBIR in medical applications [12]. For example, the Pittsburgh Supercomputing Center developed a system that 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* systemthat 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 [acute lymphocytic leukemia (ALL) and acute myelogenous leukemia (AML)].

The remainder of the paper is organized as follows: Section II 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 III. Section IV provides the experimental results and Section V 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 [distributed telemicroscopy (DT)], an intelligent archival (IA) subsystem, and an image-guided decision support (IGDS) subsystem.

The DT subsystem of PathMiner enables primary users to communicate remotely and share information. The DT subsystem features autofocusing, shared graphics, and text messaging. It is a crucial component for physicians to exchange diagnostic opinions. The IA subsystem performs automatic and remote control of the microscope. The 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 Fig. 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.

Fig. 2 shows the system architecture when the IA subsystem, IGDS subsystem, and the GCM module are pipelined to perform unsupervised specimen analysis. Subsystems can be run on single machine or distributed nodes using transmission control protocol/Internet protocol as the communication protocol. In Fig. 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).

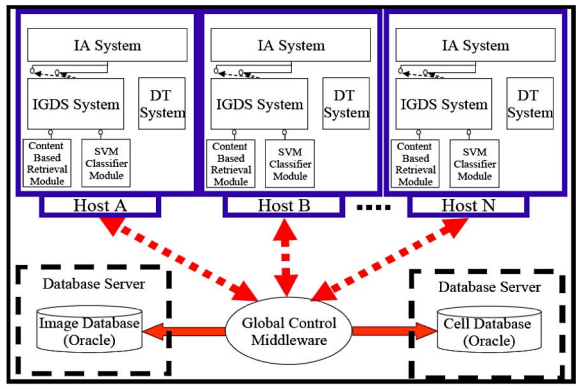

Fig. 1. System architecture of the PathMiner system.

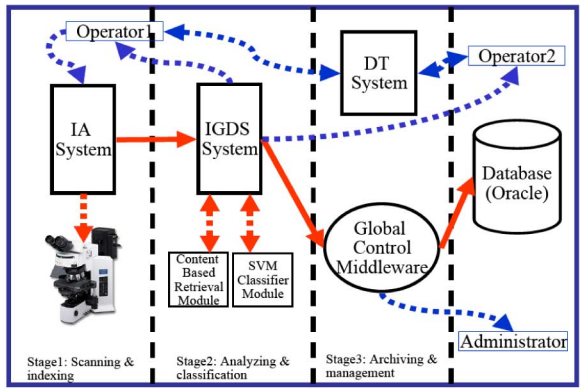

Fig. 2. Pipeline structure of the automatic decision support procedure for classifying the digitized blood cells. The IA system is used to control a microscope to automatically capture the cell images. The IGDS is the core component for performing image analysis and content-based image retrieval. The Global Control Middleware (GCM) is designed to enable the system administrator to adjust the workload and the authority of the users. The DT is used for communication among users.

A. IA System

Prior to the development of the 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 that 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 that 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.

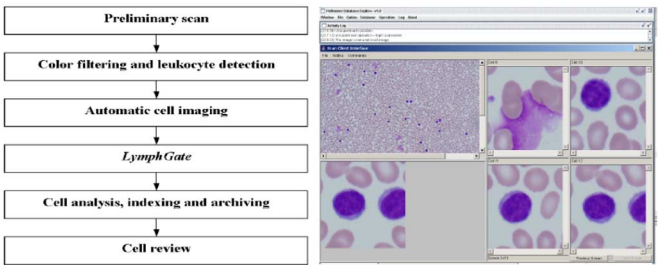

Fig. 3. Workflow (left panel) and the user interface (right panel) of the IA system. The prescanned result of an image patch under 10× magnification is on the left-top subplot of the right panel.

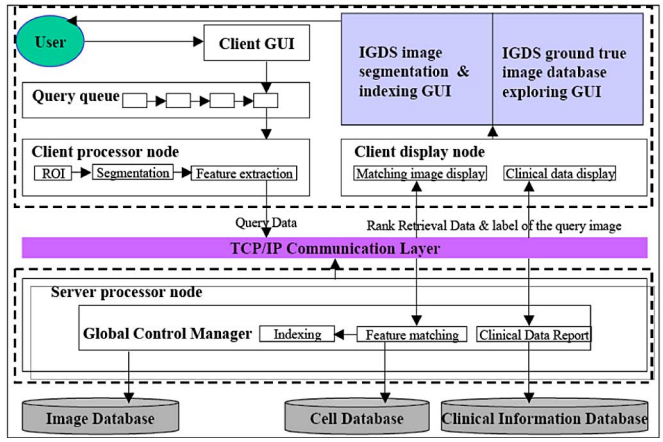

Fig. 4. System architecture of the (IGDS) subsystem in PathMiner.

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 workflow and the IA user interface are illustrated in Fig. 3. The CAM module first directs the robotic scope to perform an unsupervised pilot scan over the specimen at low resolution (10×), 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 cell 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

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

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 that 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 MB of RAM.

B. IIGDS System and GCM

The overall system architecture of the IGDS subsystem is shown in Fig. 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 Fig. 2, a preselected 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 that 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, namely, color as defined in Luv color space, nuclear and cytoplasmic area, the cytoplasm/nuclear ratio, and the texture measurements. The image metrics that 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 is 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] that 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 humans perceive as being discriminative among textures. The computational model for textons, introduced by Leung and Malik [16], depicts them as cluster centers in a feature space that is generated in response 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 III.

As the query is generated, it is sent to the IGDS system through a GCM. The feature measurements of the unknown samples are compared with the signatures that 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

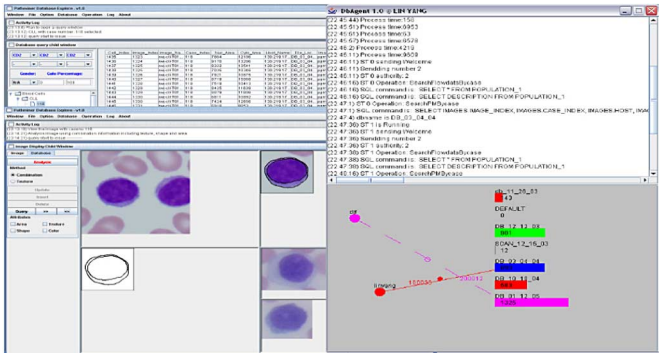

Fig. 5. Graphic user interface of the (IGDS) system is shown on the left and theGCM is shown on the right. The dotted line in the GCM represents the requirement issued from the user, the solid line denotes the requirement is fulfilled by GCM.

and protein profiles) are returned to the individual who is seeking decision support.

The 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 that 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; and 3) an administration tool that can be used to control the privileges of users. Fig. 5 shows the GUI 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. DT System

The 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 IGDS subsystem. A snapshot of the DT client interface is shown in Fig. 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.

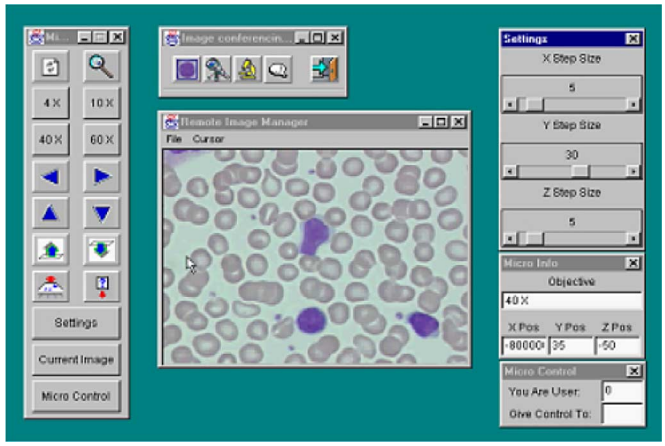

Fig. 6. GUI of the DT system.

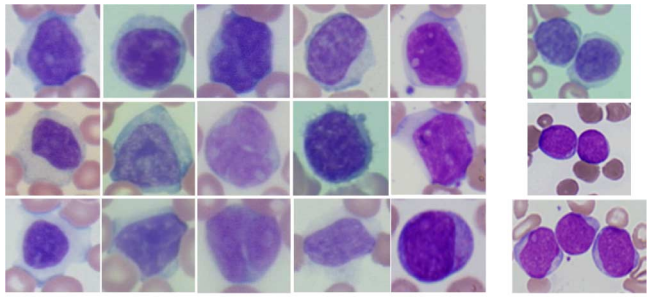

Fig. 7. (a) Representative cell images. From left to right: CLL, MCL, FCC, and acute leukemia (ALL and AML, respectively). From top to down: good, fair, and poorly representative cells. (b) Some samples of touching cells.

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 that is important for classification. An RCGVF snake [14] was developed specifically to segment the nuclei and the cytoplasm of the cells. We first apply an L_2E robust estimation [17] to provide a rough estimation of the outer boundaries of the cells inside the ROI. A 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 1 s on a Pentium PC with 1.5 GHz CPU and 512 MB memory.

Fig. 7(a) shows good, fair, and poorly representative imaged cells from each class that were stained with hematoxylin and eosin. In our testing set, which contains 3691 cell images, about 35% belong to good images, 40% to the fair images, and the rest 25% belong 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 of 90.1% on the entire database. Some segmentation results are shown in Fig. 8. For more details about the color gradient and RCGVF snake, we refer readers to [14].

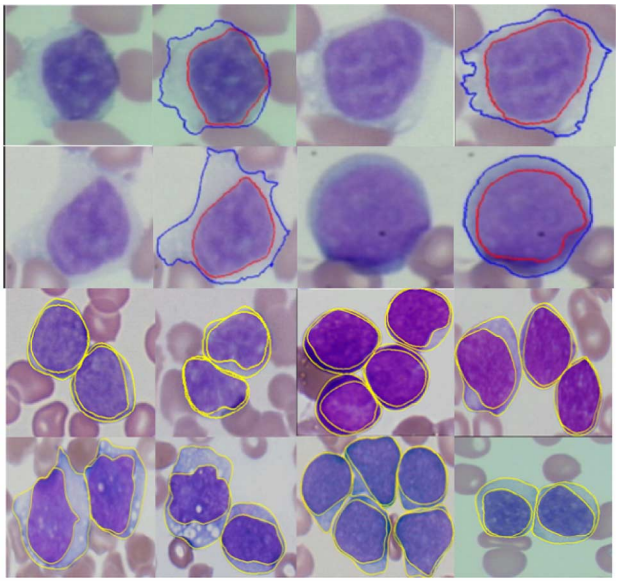

Fig. 8. Segmentation results on weak boundary cases and touching cell cases.

TABLE I
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

The segmentation algorithm is further extended to manage touching cells segmentation [20] and tested on a dataset that contains 207 touching cells [shown in Fig. 7(b)]. In the touching cell dataset, we obtained an average segmentation accuracy of 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 show 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 asses 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; and 3) the close contour reconstructed from EFD is always closed.

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

$$a_n = \frac{S}{2n^2\pi^2}\sum_{i=1}^{M}\frac{\Delta x_i}{\Delta s_i}\left[\cos\frac{2n\pi s_i}{S} - \cos\frac{2n\pi s_{i-1}}{S}\right]$$

$$b_n = \frac{S}{2n^2\pi^2}\sum_{i=1}^{M}\frac{\Delta x_i}{\Delta s_i}\left[\sin\frac{2n\pi s_i}{S} - \sin\frac{2n\pi s_{i-1}}{S}\right]$$

$$c_n = \frac{S}{2n^2\pi^2}\sum_{i=1}^{M}\frac{\Delta y_i}{\Delta s_i}\left[\cos\frac{2n\pi s_i}{S} - \cos\frac{2n\pi s_{i-1}}{S}\right]$$

$$d_n = \frac{S}{2n^2\pi^2}\sum_{i=1}^{M}\frac{\Delta y_i}{\Delta s_i}\left[\sin\frac{2n\pi s_i}{S} - \sin\frac{2n\pi s_{i-1}}{S}\right] \quad (2)$$

where $s_i, S = \sum_{j=1}^{i(M)}\Delta s_j, \Delta s_i = \sqrt{(\Delta x_i)^2 + (\Delta y_i)^2}, \Delta x_i = (x_i - x_{i-1})$, and $\Delta y_i = (y_i - y_{i-1})$. The Δx_i and Δy_i are the changes in the x and y projection of the chain code at the ith 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 weight of each feature, and f_i is the ith 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 that 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 [16], [25]–[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 0 and standard deviation was 1. Since the images were acquired under different

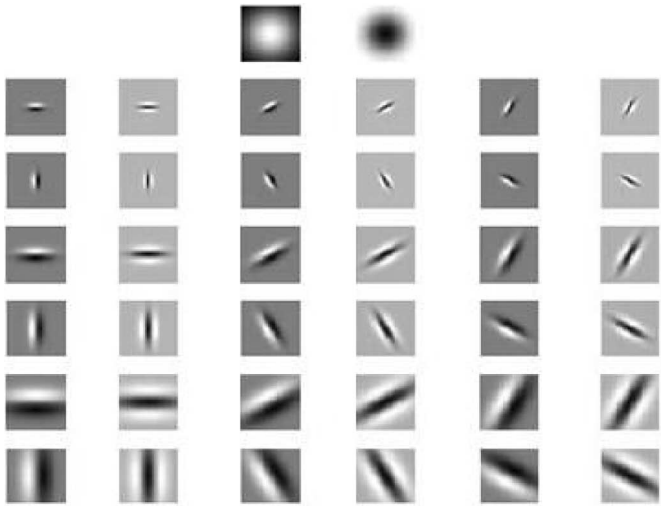

Fig. 9. M8 filter bank.

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 Fig. 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 three scales (scale values) = (1,3), (2,6), (4,12) and a bar filter at the same three 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 8-D feature vector.

The textures 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 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 histogram. 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

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

where I denotes the cell image, i is the ith 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 sizes 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 occurrences of the texton in the image.

2) Classification: The SVM was first introduced in [28] for binary classification problems. 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

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

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* that minimizes

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

where ξ_i are called slack variables that 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 $\mathbf{x}_i$ denote the feature vector, and y_i represent the ground true labels. We minimize (6) by maximizing the dual problem of (6) which involves 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

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

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

In order to extend the method to multiclass problems, 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.

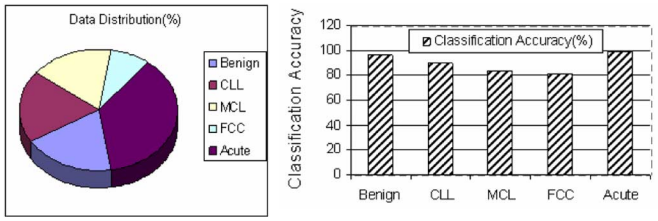

Fig. 10. Data distribution and the classification results using one-against-one SVM and tenfold validation on the first dataset that contains 3691 testing images.

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-in charge-coupled device 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 that 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 tenfold cross-validation result was reported. The second dataset, referred to as the open dataset, contained 1200 imaged cells exhibiting large variations in staining characteristics that were used only for testing.

The cell types in both datasets included a mixed set of MCL, CLL, FCC, acute leukemia, and benign. The imaged cells were collected from the Hospital of the University of Pennsylvania, Philadelphia, 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, nine 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 werereported based on a tenfold cross validation. The result is listed in Fig. 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 tenfold 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 comparision, we also performed tenfold 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

TABLE II
CONFUSION MATRIX (TENFOLD 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
CONFUSION MATRIX (TENFOLD CROSS VALIDATION) USING THE ONE-AGAINST-ONE SVM

	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

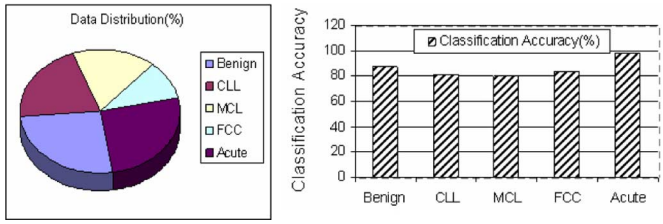

Fig. 11. Data distribution and classification results on the open dataset that contains 1200 testing images. The SVM classifiers are trained using the closed dataset and never retrained for open-dataset test.

better results 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 five MCL, six CLL, three FCC, eight acute leukemia, and 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 classification results are shown in Fig. 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 interclass similarities and intraclass variations that were never seen during the training.

In Fig. 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, whereas the last one is misclassified as CLL. The second row represents CLL whereas 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 Fig. 12, we can see that there exist interclass similarities and intraclass variations thus making the multiclass cell classification a quite challenging problem.

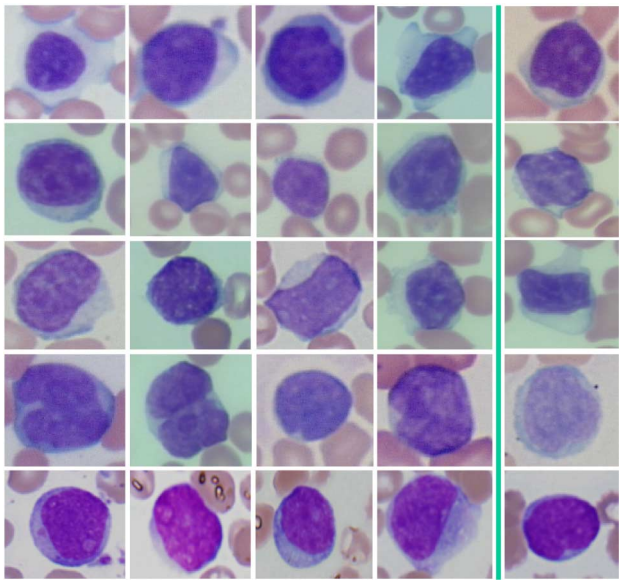

Fig. 12. Multiclass classification results using one-against-one SVM. The left four columns are correct classified samples, and the right fifth column shows the failed images.

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 workload 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 analyses. 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 multicell 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].

Acknowledgment

University of Medicine and Dentistry of New Jersey wishes to thank and acknowledge IBM for providing free computational power and technical support for this research through the World Community Grid.

References

[1] O. Debeir, C. Decaestecker, J. L. Pasteels, I. Salmon, R. Kiss, and H. P. Van, "Computer-assisted analysis of epiluminenescence microscopy images of pigmented skin lesions," *Cytometry*, vol. 37, no. 4, pp. 255–266, 1999.

[2] M. Cenci, C. Nagar, and A. Vecchione, "PAPNET-assisted primary screening of conventional cervical smears," *Anticancer Res.*, vol. 20, no. 5, pp. 3887–3889, 2000.

[3] M. R. Kok, Y. T. van Der Schouw, M. E. Boon, D. E. Grobbee, L. P. Kok, P. G. Schreiner-Kok, Y. van der Graaf, H. Doornewaard, and J. G. van den Tweel, "Neural network-based screening (NNS) in cervical cytology: No need for the light microscope?," *Diagn. Cytopathol.*, vol. 24, no. 6, pp. 426–434, 2001.

[4] J. Smith, J. Svirbely, and C. Evans, "RED: A red-cell antibody identification system," *J. Med. Syst.*, vol. 9, pp. 121–137, 1985.

[5] D. R. Thursh, F. Marby, and A. Levy, "Computers and video discs in pathology education: ECLIPS as an example of one approach," *Hum. Pathol.*, vol. 17, no. 1, pp. 216–218, 1986.

[6] B. N. Nathwani, K. Clarke, T. Lincoln, C. Berard, and C. Taylor, "Computers and video discs in pathology education: ECLIPS as an example of one approach," *Hum. Pathol.*, vol. 28, no. 9, pp. 117–121, 1997.

[7] C. Faloutsos, W. Equitz, M. Flickner, W. Niblack, D. Petkovic, and R. Barber, "Efficient and effective querying by image content," *J. Intell. Inf. Syst. Integr. Artif. Intell. Database Technol.*, vol. 3, no. 3, pp. 231–262, 1994.

[8] A. Pentland, R. W. Picard, and S. Scarloff, "Photobook: Content based manipulation of image databases," *Int. J. Comput. Vis.*, vol. 18, pp. 233–254, 1996.

[9] J. Z. Wang, G. Wiederhold, O. Firschein, and X. W. Sha, "Content-based image indexing and searching using Daubechies wavelets," *Int. J. Digital Libr.*, vol. 1, no. 4, pp. 311–328, 1998.

[10] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 1027–1037, Aug. 2002.

[11] J. Li, J. Z. Wang, and G. Widerhold, "IRM: Integrated region matching for image retrieval," *Proc. ACM Multimedia*, vol. 1, no. 4, pp. 147–156, 2000.

[12] F. Schnnorrenberg, C. S. Pattichis, C. N. Schizas, and K. Kyriacou, "Content-based retrieval of breast cancer biopsy slides," *Technol. Health Care.*, vol. 8, no. 5, pp. 291–297, 2000.

[13] A. W. Wetzel, "Computational aspects of pathology image classification and retrieval," *J. Supercomput.*, vol. 11, no. 1, pp. 279–293, 1997.

[14] L. Yang, P. Meer, and D. Foran, "Unsupervised segmentation based on robust estimation and color active contour models," *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 3, pp. 475–486, Sep. 2005.

[15] L. Yang, W. Chen, P. Meer, G. Salaru, M. D. Feldman, and D. J. Foran, "High throughput analysis of breast cancer specimens on the grid," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2007, pp. 617–625.

[16] T. Leung and J. Malik, "Recognizing surfaces using three-dimensional textons," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1999, pp. 1010–1017.

[17] D. W. Scott, "Parametric statistical modeling by minimum integrated square error," *Technometrics*, vol. 43, pp. 274–285, 2001.

[18] C. Xu and J. L. Prince, "Snakes, shapes and gradient vector flow," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 359–369, Mar. 1998.

[19] G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd ed. New York: Wiley, 1982.

[20] L. Yang, O. Tuzel, P. Meer, and D. J. Foran, "Touching cells segmentation in hematologic specimens using concave vertex graph," presented at the Int. Conf. Med. Image Comput. Comput. Assist. Interv., New York, 2008.

[21] A. N. Moga and M. Gabbouj, "Parallel marker-based image segmentation with watershed transformation," *J. Parallel Distrib. Comput.*, vol. 51, no. 1, pp. 27–45, 1998.

[22] P. S. U. Adiga and B. B. Chaudhuri, "An efficient method based on watershed and rule-based merging for segmentation of 3D histo-pathological images," *Pattern Recognit.*, vol. 34, no. 7, pp. 1449–1458, 2001.

[23] D. Comaniciu, P. Meer, and D. Foran, "Image-guided decision support system for pathology," *Mach. Vis. Appl.*, vol. 11, pp. 213–224, 1999.

[24] F. P. Kuhl and C. R. Giardina, "Elliptic Fourier features of a closed contour," *Comput. Graph. Image Process.*, vol. 18, pp. 236–258, 1982.

[25] O. Cula and K. Dana, "3D texture recognition using bidirectional feature histograms," *Int. J. Comput. Vis.*, vol. 59, no. 1, 2004.

[26] D. Heeger and J. Bergen, "Pyramid-based texture analysis/synthesis," in *Proc. ACM Int. Conf. Exhib. Comput. Graph. Interact. Tech.*, 1995, pp. 229–238.

[27] M. Varma and A. Zisserman, "Classifying images of materials: Achieving viewpoint and illumination independence," in *Proc. Eur. Conf. Comput. Vis.*, 2002, vol. 3, pp. 255–271.
[28] C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learn.*, vol. 20, pp. 1–25, 1995.
[29] C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
[30] O. Tuzel, L. Yang, P. Meer, and D. J. Foran, "Classification of hematologic malignancies using texton signatures," *Pattern Anal. Appl.*, vol. 10, pp. 277–290, 2007.

Lin Yang (S'09) received the B.S. degree in electrical and computer engineering from Xian Jiaotong University, Xian, Shaanxi, P.R. China, in 1999, and the M.S. degree from the Image Processing Center, Xian Jiaotong University, in 2002, and the another M.S. and Ph.D. degrees in electrical and computer engineering from Rutgers University, New Brunswick, NJ, in 2006 and 2009, respectively.

He is currently an Assistant Professor with the Department of Radiology, University of Medicine and Dentistry of New Jersey (UMDNJ)—Robert Wood Johnson Medical School, Piscataway, NJ, where since 2009, he has been with the faculty in the Department of Radiology. Since 2009, he has been also with the Center for Biomedical Imaging and Informatics (CBII) as a Senior Research Staff Member. His current research interests include different areas of medical imaging, computer vision, and machine learning and also working on the design and development of content-based image and video retrieval and 2-D/3-D medical image analysis including tumor detection, segmentation, registration, and tracking.

Oncel Tuzel (S'04–M'09) received the B.S. and M.S. degrees in computer engineering from Middle East Technical University, Ankara, Turkey, in 1999 and 2002, respectively, and the Ph.D. degree in computer science from Rutgers University, Piscataway, NJ in 2008.

He is currently a Research Scientist at Mitsubishi Electric Research Laboratories, Cambridge, MA. He is the author or coauthor of more than 15 papers in the filed of computer vision. He is the holder of several patents, awarded or pending. His current research interests include computer vision, computer graphics, machine learning, and statistical pattern recognition. Dr. Tuzel is a Student Member of the IEEE Computer Society. He was on the program committee of several international conferences including Computer Vision and Patern Recognition. He was the recipient of the Best Paper Runner-Up Award at the IEEE Computer Vision and Pattern Recognition Conference in 2007.

Wenjin Chen received the Bachelor of Medicine degree from Beijing Medical University (now Peking University Health Science Center), Beijing, China, in 1997, and the Ph.D. degree in computational molecular biology from the University of Medicine and Dentistry of New Jersey (UMDNJ) and Rutgers University, Piscataway, in 2005.

She is currently serving as Associate Director, Biomedical Imaging at the Center for Biomedical Imaging and Informatics, UMDNJ. Her current research interests include tissue microarray analysis, digital microscopy, image pattern recognition, and medical informatics.

Peter Meer (S'84–M'86–SM'95) received the Dipl. Eng. degree from Bucharest Polytechnic Institute, Bucharest, Romania, in 1971, and the D.Sc. degree from Technion, Israel Institute of Technology, Haifa, Israel, in 1986, both in electrical engineering.

From 1971 to 1979, he was with the Computer Research Institute, Cluj, Romania, where he was enagaged in research on digital hardware. Between 1986 and 1990, he was an Assistant Research Scientist at the Center for Automation Research, University of Maryland at College Park, College Park. In 1991, he joined the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, and is currently a Professor. He has held visiting appointments in Japan, Korea, Sweden, Israel and France. His current research interests include application of modern statistical methods to image understanding problems.

Prof. Meer was on the organizing committees of numerous international workshops and conferences. He was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE between 1998 and 2002, is a member of the Editorial Board of *Pattern Recognition*, and was a Guest Editor of *Computer Vision and Image Understanding* for a special issue on "Robust Statistical Techniques in Image Understanding." He is the coauthor of an award winning paper in *Pattern Recognition* in 1989, the Best Student Paper in 1999, and the Best Paper in the 2000 IEEE Conference on Computer Vision and Pattern Recognition.

Gratian Salaru received the M.D. degree from the University of Medicine and Pharmacy, Timisoara, Romania, in 1995. After three years of forensic pathology at the °Mina Minovici± Institute of Forensic Medicine, Timisoara, he completed an additional five years of training and completed the Pathology Residency Program in 2004 at Robert Wood Johnson Medical School (RWJMS)/University of Medicine and Dentistry of New Jersey (UMDNJ), New Brunswick, from where he also completed a Fellowship in Hematopathology in 2006.

He is Board Certified in Anatomical and Clinical Pathology and is currently an Assistant Professor of Pathology and Laboratory Medicine at RWJMSl/UMDNJ. His current research interests include areas of clinical pathology, digital microscopy, medical informatics, and hematopathology.

Lauri A. Goodell received the M.D. degree from the University of Medicine and Dentistry of New Jersey (UMDNJ)—Robert Wood Johnson Medical School (RWJMS), New Brunswick, in 1991.

She is currently an Associate Professor and Director of Hematopathology in the Department of Pathology and Laboratory Medicine, UMDNJ—RWJMS, as well as the Director of the Immunohistochemistry Core Research Laboratory, The Cancer Institute of New Jersey, New Brunswick. Her current research interests include pathophysiology of hematopoietic diseases, proteomics, image pattern recognition, and telepathology.

David J. Foran (S'89–M'91) received the B.S. degree from Rutgers University, New Brunswick, NJ, in 1983 and the Ph.D. degree in biomedical engineering from the University of Medicine and Dentistry of New Jersey (UMDNJ) and Rutgers University, Piscataway, NJ, in 1992.

From 1984 to 1985, he served as a Physics Instructor at New Jersey Institute of Technology, Newark, NJ. From 1986 to 1988, he worked as a Junior Scientist at Johnson & Johnson Research, Inc., North Brunswick, NJ. He received one year of postdoctoral training at the Department of Biochemistry, UMDNJ—Robert Wood Johnson Medical School (RWJMS) in 1993. He joined the faculty at RWJMS in 1994, where he is currently a Professor of pathology and radiology and the Director of the interdepartmental Center for Biomedical Imaging and Informatics. He also serves as the Associate Director for research for the university-wide Informatics Institute. He is a member of the Graduate Faculty in the Program in Computational Molecular Biology, Rutgers University. His research interests include quantitative, biomedical imaging, computer-assisted diagnostics, and medical informatics.