Image Mining for Investigative Pathology Using Optimized Feature Extraction and Data Fusion

Wenjin Chen[1], Peter Meer[3], Bogdan Georgescu[4], Wei He[1],Lauri A. Goodell[2], David J. Foran[1, 2]

[1]Center for Biomedical Imaging & Informatics, University of Medicine & Dentistry of New Jersey, Piscataway, NJ 08854

[2]Department of Pathology & Laboratory Medicine, University of Medicine & Dentistry of New Jersey, Piscataway, NJ 08854

[3] Department of Electrical and Computer Engineering, Center for Advanced Information Processing, Rutgers University, Piscataway, NJ 08854

[4]Siemens Corporate Research, Integrated Data Systems Department, Princeton, NJ

For correspondence please contact Wenjin Chen at:Room R203, 675 Hoes Lane, Piscataway, NJ 08854, Telephone: 732-235-5680, Fax: 732-235-4825, Email: wjc@pleiad.umdnj.edu

Summary

In many subspecialties of pathology, the intrinsic complexity of rendering accurate diagnostic decisions is compounded by a lack of definitive criteria for detecting and characterizing diseases and their corresponding histological features. In some cases, there exists a striking disparity between the diagnoses rendered by recognized authorities and those provided by non-experts. We previously reported the development of an *Image Guided Decision Support* (IGDS) system, which was shown to reliably discriminate among malignant lymphomas and leukemia that are sometimes confused with one another during routine microscopic evaluation. As an extension of those efforts, we report here a web-based intelligent archiving subsystem that can automatically detect, image and index new cells into distributed ground-truth databases. Systematic experiments showed that through the use of robust texture descriptors and density estimation based fusion the reliability and performance of the governing classifications of the system were improved significantly while simultaneously reducing the dimensionality of the feature space.

Keywords

Automated digital microscopy; Unsupervised cell imaging; Content-based image retrieval; Texture analysis; Data fusion.

1 Introduction:

1.1 Clinical Significance

A differential diagnosis provides the basis for how patients are treated, which medications are appropriate, and what levels of risk are justified. As new treatments and therapies become available it is essential to distinguish among subclasses of pathologies [1]. The peripheral blood of patients is routinely screened for abnormalities, however, the subtle visible differences exhibited by some disorders can lead to a significant number of false negatives during routine microscopic evaluation of specimens.

Mantle cell lymphoma (MCL) is an intermediate grade lymphoma with a 3-5 year median survival rate [2, 3, 4]. MCL morphology is generally described as a monotonous proliferation of small to medium sized lymphoid cells with scant cytoplasm, variably irregular, round or indented nuclei, dispersed chromatin, and inconspicuous nucleoli. However, the disease may exhibit a spectrum of presentations, which can sometimes be confused with other lymphomas that follow a less aggressive clinical course. Chronic Lymphocytic leukemia (CLL) and Follicular Center Cell lymphoma (FCC) are two examples which were considered in our studies. The main objective of our research was to develop and optimize image-based methods which would provide quick, reliable decision support in detecting and characterizing these disorders.

Immunophenotyping with flowcytometry is considered a definitive approach for reliably differentiating among lymphoproliferative diseases [3, 5, 6, 7]. However, because of the time and expense of implementing studies, it is not generally utilized unless a case is first categorized as suspicious during microscopic evaluation. Throughout the course of our studies the diagnosis determined by immunophenotyping was considered as the gold-standard for gauging the performance of the image guided approach [8].

1.2 Technical Significance

Recent literature ascribes much of the difficulty in rendering consistent diagnoses to the subjective impressions of observers and shows that, when morphologic cell classification is based upon computer-aided analysis, objectivity and reproducibility can be made to improve considerably [9, 10, 11, 12]. Using these techniques it may be possible to detect and track subtle changes in measurable parameters leading to the discovery of novel diagnostic clues, which may not be apparent by human visual inspection alone.

Developing approaches that can reliably transform complex diagnostic concepts into well-defined algorithmic procedures is an active area of research [13, 14, 15, 16, 17]. Diagnostic pathology offers a rich environment for conducting such studies. Thus, several major projects in artificial intelligence have focused on pathology. These include the *Pathex* framework and the *Pathex/Red* system [18] for assisting pathologists with laboratory data at Ohio State University, *ECLIPS* [19] at the University of Illinois Urbana, as well as the *PathFinder* project on anatomic pathology diagnosis [20] at the University of Southern California and Stanford. In *PathFinder* an expert system provides a differential diagnosis based on the initial histological feature(s) observed by the pathologists, and suggests to the user additional histological features for observation that are likely to narrow the differential diagnosis, thus helping to screen for observations which are incompatible with a given disease or disorder.

While the mechanisms for content-based access to alphanumeric data have been extensively studied, and are now considered relatively well understood, content-based image access remains elusive and is an active area of research in the computer vision

and image processing communities, as well as in disparate application domains, such as remote sensing, diagnostic medicine, molecular biology, pharmacy, and computer aided design. Technologies that capture, describe, and index the visual essence of multimedia objects rely on the methods and principles of image analysis, pattern recognition, and database theory. This relatively new area of research spans a spectrum of applications [21, 22, 23, 24]. A survey of content-based image retrieval (CBIR) was recently presented by Antani [25].

Several general-purpose CBIR systems have been developed such as the IBM *QBIC* System [26], the *Photobook* System [27], the *WBIIS* System [28], the *Blobworld* System [29] and the *SIMPLIcity* System [30]. These systems are not suited for most pathology applications, however, because of the special characteristics of digitized pathology such as the high resolution of the images and chromatic profiles of the stained specimens. Over the past few years there has been increased interest and effort applied to utilizing CBIR in medical applications [31, 32, 33]. Individual strategies and approaches differ according to the degree of generality (general purpose versus domain specific), level of feature abstraction (primitive feature versus logical features), overall dissimilarity measure used in retrieval ranking, database indexing procedure, level of user intervention (with or without relevance feedback), and by the methods used to evaluate their performance. The Pittsburgh Supercomputing Center has developed a system which utilizes global characteristics of images to provide a measure of Gleason grade of prostate tumors [34]. Wang from Pennsylvania State University emphasizes the use of wavelet technology and Integrated Region Matching (IRM) distances for characterizing pathology images [35]. The system indexes block segments of images at different scales by partitioning the original image into smaller overlapping blocks. The CBIR engine is interfaced with a server that allows users to browse portions of the original matched image at different scales. This system differs from the system that we are developing, both in design and purpose. The system from Penn State emphasizes the general tasks of retrieval and browsing whereas the system that we have proposed focuses on developing a set of portable, web-based tools for collaborative studies and clinical decision support.

1.3 The *PathMiner* Project

We have already reported the design and development of an *Image Guided Decision Support* (IGDS) system, which was shown to reliably discriminate among a set of lymphoproliferative disorders that can sometimes be confused with one another during routine microscopic evaluation. The IGDS system automatically identifies and retrieves images, diagnosis, and correlated clinical data of those cases from within a "gold standard" database whose spectral and spatial profiles are most similar to a given query image. The system suggests the most likely diagnosis based on majority logic of the retrieved cases. Man-machine performance comparison studies showed that the image-guided approach provided significant improvement in discriminating among disorders while simultaneously reducing the frequency of false negatives. One project which is closely related to our research is the *PathMaster* system from Yale University, which is a content-based retrieval system that was used to discriminate among specimens of Mantle Cell Lymphoma (MCL) and small cell lymphomas [36] that are prepared using a touch prep protocol. This system implements semi-automatic segmentation utilizing commercial, off-the-shelf image segmentation (Adobe PhotoShop) and does not exploit the potential of advanced computer vision techniques. Within this system each cell can be represented by more than 2000 features, but the question of which of the features belong to an optimal subset was not explored.

The *PathMiner* project, that we have undertaken, started as an effort to establish a large-scale web-based pathology image database, which could provide clinical decision support and medical education based on statistical pattern recognition and CBIR. Since each of the diseases under study may exhibit a spectrum of morphologies, it was important that the database contain a sufficiently large number of cases. This allows us to effectively exploit the statistical methods that are utilized for discriminating among disorders.

The *PathMiner* project consists of three subsystems: the *Distributed Telemicroscopy* (DT) subsystem, the *Intelligent Archiving* (IA) subsystem, and the IGDS subsystem. The first generation DT subsystem [6] provides a distributed telepathology environment by enabling multiple users from disparate clinical and research sites to simultaneously control robotic microscopes from remote locations while each session participant receives a digital broadcast of the specimen. The IA subsystem facilitates the automated population and management of databases through the use of unsupervised scanning and indexing of candidate lymphocytes. In this paper, we report how through the use of robust texture descriptors and density estimation based data fusion the reliability and performance of the governing classifications was significantly improved while simultaneously reducing the dimensionality of the feature space.

With minimal exceptions, the software modules of *PathMiner* project were implemented using the JAVA programming language to provide maximum portability.

2 Methods

2.1 Classification Optimization of IGDS

2.1.1 Optimization Condition

Cross validation methods are widely used in statistical pattern recognition to determine the robustness of an algorithm by evaluating generalization performance based on "resampling". The k-fold cross validation approach first randomly divides data into k subsets of approximately the same size. The "test algorithm" is then trained with k-1 subsets, and tested with the remaining subset to generate the appropriate performance measure. The training and testing session is executed k times, with each iteration utilizing different subsets for testing. Therefore, when smaller k's are used fewer samples are utilized for training thus creating a heightened challenge for the algorithm's capacity to produce the same performance measure.

Because of limited sample size, a ten-fold cross validation was used in the original IGDS study [6] . In order to further evaluate the robustness and stability of the IGDS system, a balanced mix of over 900 test cells was recently established. Therefore, experiments reported in the current study used the much more stringent two-fold cross validation and were aimed to maximize correct classification rate, defined as

$$correct\,classification\,rate = \frac{number\,of\,correctly\,classified\,cells}{total\,number\,of\,cells} \quad (1)$$

2.1.1 Algorithms Used in the Current IGDS

The original IGDS scheme used a Simplex strategy [37] based on an eight nearest neighbor search [38] across the weighted-sum of distances. Weighting factors for shape, texture, and area, were tested on the new dataset. In order to gain better insight into the nature of the high-dimensional data set, additional experiments were conducted using joint-ordering (sum of rank) [6, 39]. These studies demonstrated a much more

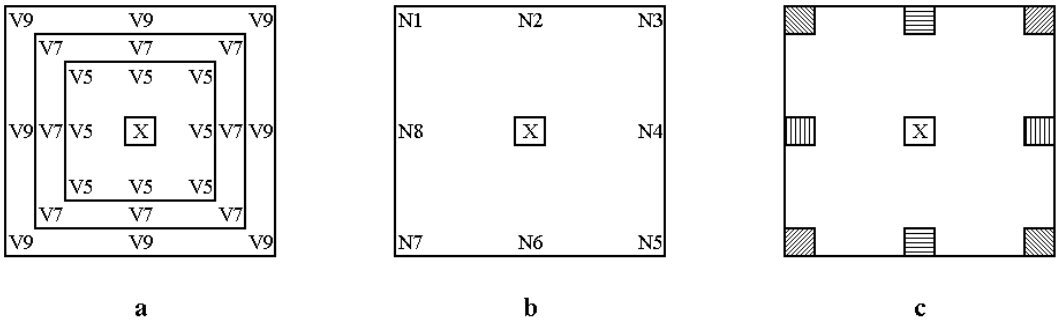

Figure 1. Neighborhoods used in MRSAR and RSAR methods. a: Neighborhood used in the MRSAR method. b: Non-symmetrical neighborhood used in the RSAR method. c: Symmetrical neighborhood used in the RSAR method.

stable performance and revealed that area and texture feature measurements contributed most to an optimized classification. These observations led us to focus our efforts on investigating the potentials of multivariate data fusion for improving the discrimination performance and robustness of the governing algorithms.

2.1.3 Region Simultaneous Autoregressive (RSAR) model

The simultaneous autoregressive (SAR) model is a linear regressive model often used in imaging applications involving texture analysis. It defines the texture feature by computing SAR coefficients θ, the intersect μ, and the standard deviation of the residual ε over a specified area as

$$I(x) = \mu + \sum_{y \in N} \theta(y) I(y) + \varepsilon(x) \qquad (2)$$

in neighborhood N. Depending on data normalization, μ can sometimes be considered 0 and omitted from the regression calculation.

SAR regression is usually performed over overlapping neighborhoods centered at the pixel of interest. To achieve adequate texture analysis, the size of the overlapping neighborhoods must be sufficiently large to capture the texture characteristics, while remaining small enough to maintain low texture variability across the image. The Multi-resolution SAR (MRSAR) uses multiple neighborhood sizes to accommodate complex texture types.

In the first generation IGDS prototype, the MRSAR descriptor was computed for each region-of-interest (ROI) as described by Mao and Jain [40]. Three different neighborhood sizes (5x5, 7x7 and 9x9, see Figure 1a) were simultaneously investigated to provide multi-scale texture information. At each examined nuclear pixel, the descriptor was estimated using pixel luminance (L^*) value throughout a 21x21 window. By combining symmetric neighbor pixels in the estimation, the resulting texture feature vector was 15 dimensional, (4 neighbor directions + 1 residual) x 3 neighborhood sizes. The least square method was used in the linear regression. A fifteen dimensional mean vector and 15x15 covariance matrix of the texture feature were then computed for the entire ROI and stored for database query.

In an attempt to improve the robustness of the estimation while reducing the dimensionality of the texture descriptor, the SAR model was modified using an approach that we refer to as Region SAR (RSAR). This algorithm differs from the MRSAR model in two aspects. First, since the original MRSAR implementation was designed with the purpose of texture segmentation, estimates are generated in relatively small local windows to maintain local information. Since the desired parameters are of high dimensionality (15), the estimation from a set of 21x21 windows is not very robust from a statistical point of view. Another limitation of the MRSAR implementation was that it positioned the 21x21 windows with considerable overlap, giving rise to a substantial amount of redundancy in computing the estimation of local texture features. Since texture-based segmentation was not the goal of our application, we reasoned that it would not be necessary to estimate local texture features in the 21x21 windows. Instead, the SAR parameters are only estimated in the ROIs, i.e. cell nuclei. This strategy proved to be more robust since the estimation was conducted with a much richer informational content, while simultaneously maintaining efficiency in terms of computational complexity.

Second, although multiple neighborhood sizes capture more *contextual* information for each pixel, the actual information that is retrieved from multiple neighborhood sizes is strongly correlated with one another and does not provide much of an advantage in terms of texture discrimination. On the contrary, it suffers from the potential drawback of the "curse of high dimensionality" [41, 42]. To investigate this logic only a single window was utilized during the course of our experiments to evaluate the new descriptor. A series of systematic experiments was conducted using varying neighborhood sizes to determine the optimal dimensions for this application (please see the **Results** section below). A set of comparative performance studies was also conducted to evaluate the use of symmetric and non-symmetric neighborhoods. In the case of the non-symmetric approach each of the eight neighbors in the window was considered separately (see Figure 1b), whereas in the case of symmetric neighborhoods, the symmetrically located pixels were combined, thus reducing dimensionality by half (see Figure 1c).

2.1.4 Data Fusion

Data fusion is an active area of research with application to the field of CBIR, which attempts to integrate the informational content of disparate sources or modalities in order to improve precision and accuracy of retrievals from an image database. Based upon our previous work in CBIR, we concentrated our efforts on feature measurements for area and texture.

Probability density based approaches are one of the methods that are being actively evaluated for use in data fusion applications. Probability density strategies are derived from the popular theory that the joint conditional probability of several independent events is the product of their individual conditional probabilities. The essence of effectively using a probability-based method in data fusion is dependent upon appropriate estimation of conditional probability densities.

For the purposes of analysis a set of governing equations was established. Letting $\underline{i}$ serve as the class indicator, where $\underline{i}$ = 1 for Mantle Cell Lymphoma (MCL), 2 for Chronic Lymphocytic Leukemia (CLL), 3 for Follicular Center Cell lymphoma (FCC), and 4 for Normal; P_i is the a priori probability for each class; the combined probability for each cell belonging to a certain class $\underline{i}$ is computed as

$$p_i = P_i \times \prod_{f \in features} p_i^f \quad i = 1,2,3,4 \quad features = \{area, texture\} \tag{3}$$

The proposed classification of each cell was computed as

$$class = \{i \mid p_i = \max(p_i) \quad i = 1,2,3,4\} \tag{4}$$

As in most real world scenarios, the probability density distribution is unknown. To address this problem, kernel based probability density estimation methods were used to give a robust estimation of the local probability density based on the available data. Kernel based probability density estimation methods can be understood as follows: in order to estimate the probability density at $\underline{y}$, we analyze from given data all points which fall within an ellipsoid that is centered at $\underline{y}$, while the contribution of each of these points to the estimation is based on a kernel function $\underline{k}$. It is important to generate the appropriate kernel size $\underline{h}$ for adequate estimation.
The probability density for nuclear area was estimated using

$$\hat{p}_i^{area}(y) = \frac{1}{n_i h_i} \sum_{j=1}^{n_i} k\left(\frac{|y - y_j|}{h_i}\right) \tag{5}$$

where the kernel was computed as

$$k(u) = \begin{cases} \frac{35}{32}\left(1 - \left(1 - u^2\right)^3\right) & 0 \le |u| \le 1 \\ 0 & |u| > 1 \end{cases} \tag{6}$$

with the bandwidth estimated by

$$h_i = \underset{j=1}{\overset{n_i}{median}} \left| y_j - \underset{k=1}{\overset{n_i}{median}} |y_k| \right| \quad i = 1,2,3,4. \tag{7}$$

The probability density for the RSAR texture descriptor was derived from a previous implementation [43] and in our studies was computed as,

$$\hat{p}_i^{texture}(y) = \frac{C_{k,d}}{n_i \left(\chi^2_{\gamma,p}\right)^{p/2}} \sum_{j=1}^{n_i} \left[\det\left[\frac{C_y}{t}\right]^{-\frac{1}{2}} k\left(\frac{1}{\chi^2_{\gamma,d}} D\left(y, y_j, \frac{C_y}{t}\right)\right)\right] \tag{8}$$

where D is Mahalanobis distance, and the kernel is

$$k(u) = \begin{cases} 1 - u & 0 \le |u| \le 1 \\ 0 & |u| > 1 \end{cases} \tag{9}$$

and bandwidth matrices are

$$H_y = \frac{\chi^2_{\gamma,p} C_y}{t} \tag{10}$$

where t is used to justify the kernel size since the covariance matrices C_y are singular in our case, and $\chi^2_{\gamma,p}$ is the chi-square value for $\underline{p}$ degrees of freedom and level of confidence γ ($\gamma = 0.85$ is used in these experiments).

2 Intelligent Archiving (IA) system

Prior to the development and implementation of the IA subsystem, indexing new cases into the database involved multiple steps. First, pathologists systematically reviewed specimens manually at low magnification to identify cells of interest. Each of those cells was interactively brought into focus at high magnification while capturing a digital image. Subsequently, one by one, each imaged cell was loaded into the IGDS system, automatically segmented and indexed into the image repository while an entry was created in the database with regard to the location of the corresponding images and feature measurements. The newly developed IA subsystem now serves to reduce the level of intervention on the part of pathologists since the driving software directs the robotics and imaging devices to automatically scan specimens while detecting and imaging candidate lymphocytes and extracting their image feature measurements. The IA subsystem automatically populates and manages the appropriate databases. Since the IA subsystem is web-based, it is being used to facilitate inter-institutional collaborations on the *PathMiner* project.

2.2.1 Development Platform

The original web-based, robotic microscope control module 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 Mbytes of RAM, and a Windows 2000 operating system. The software automatically images and digitizes the pathology specimens using an Olympus DC330 720-line, 3-Chip video camera and a Flashpoint 128 high-resolution frame grabber. An Olympus DP70, 12-bit color depth for each color channel, 1.45 million pixel effective resolution, single 2/3 inch CCD digital camera has recently been incorporated into the hardware configuration as an alternative high resolution imaging source.

2.2.2 Implementation

In the first generation *Distributed Telemicroscopy* (DT) prototype [6] , remote control of the robotic microscope over the web was highly interactive, requiring the operator to play a continuous, interactive role. A *Computer Assisted Microscopy* (CAM) server module was recently developed to provide intelligent control and coordination among each of the four primary devices -- the robotic stage, the motorized objective turret, an image acquisition board, and the camera -- while simultaneously performing unsupervised processing and analysis of specimens. This CAM server module features a coordinate system, which takes into account both the stage position and the optical configuration of each objective in order to accurately and reliably map coordinates between the physical specimen and the optical and digital fields of view. The system automatically corrects for hardware objective co-centering error and can be easily re-calibrated to accommodate new hardware configurations.

With the added features of the CAM module, the IA subsystem now performs *unsupervised* detection, imaging, and storage of candidate lymphocytes and management of gold standard databases. The process begins with the system performing a pilot scan of the specimen at low resolution. The output of that operation is an image map, which is subsequently filtered in $L^*u^*v^*$ color space to detect leukocytes while spatial filtering is applied to eliminate non-cellular artifacts. During the course of the scanning and filtering process, the exact stage coordinates for each candidate cell are extracted and

subsequently serve to direct the robotic scope to systematically image each candidate cell at high resolution while simultaneously segmenting the image and generating the corresponding image metrics. A second level of filtering, called the *LymphGate*, is implemented to reject candidate cells whose feature profiles are inconsistent with that of a lymphocyte. The remaining imaged lymphocytes and their corresponding image-based feature metrics are indexed into the *PathMiner* ground-truth database. In the current stage of development a certified pathologist subsequently reviews all cells that have been selected by the system prior to their becoming integrated with the core ground-truth database for quality control. Technical highlights of the entire procedure are detailed in the following sections.

2.2.2.1 Sampling strategy

Scanning sessions may be performed on the entire specimen, but are more typically applied to specific subregions as directed by operators who may view the slide locally or from remote sites. Since peripheral blood smear specimens have non-uniform morphologic distributions, scanning areas are usually specified at appropriate body-to-tail regions of the smear. There also exists an option to select multiple regions within a given specimen.

2.2.2.2 Unsupervised specimen scanning and cell detection in $L^*u^*v^*$ color space

Pilot scans may be initiated using objective lenses specified by local or remote users. The system automatically performs a quick auto-focus operation which was developed based upon Shannon's entropy [44] in order to determine the optimal focal plane for the scan. Options also exist for users to manually focus the system. Upon receiving these commands, the CAM module computes appropriate step sizes, and systematically images the required number of frames in a raster pattern while stitching slightly overlapped frames into one seamless map image. During the scanning procedure, the CAM module transmits a scaled down version of frame images to the remote user, enabling that individual to monitor the scanning and stitching process. A default scan of 5 rows and 6 columns using a 10x objective covers approximately a 3mm by 3mm region of a specimen, and requires one to two minutes for processing and transmission to a remote user utilizing a fractional T1 network connection.

A color-filtering module has been coupled with connected-component analysis algorithms in order to enable the system to automatically identify leukocytes throughout the map image. The algorithms were tested on digitized, peripheral blood smears stained with Wright Giemsa stain wherein red blood cells stain pink and the nuclei of leukocytes, including lymphocytes, neutrophils, monocytes, etc. stain in shades of blue to purple.

The algorithms begin by mapping r, g, b intensity values within the image into the $L^*u^*v^*$ color space which was introduced by CIE (*Commission Internationale de l'Eclairage*) in 1976 [45]. $L^*u^*v^*$ has become a widely used color space because it closely relates to representation in human color perception, in which the L^* dimension corresponds to the luminance and the u^*v^* dimensions relate to chrominance. The color filtering module subsequently performs a polar transformation [46] on

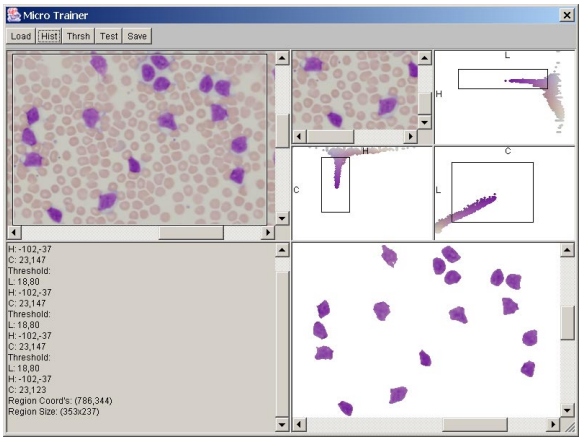

Figure 2. $L^*h^*C^*$ **based color filter training interface. Either the entire field or a subregion of the map image (upper-left) can be used to generate appropriate** $L^*h^*C^*$ **boundaries. Graphical boxes are shown in three corresponding two-dimensional plot of the three-dimensional color space (upper-right), with the subsequent filtering result shown in the lower-right quarter of the graphical interface.**

the output of these operations, resulting in $L^*h_{uv}C^*_{uv}$ color representation where h_{uv} and C^*_{uv} are computed as

$$C^*_{uv} = \left[\left(u^*\right)^2 + \left(v^*\right)^2 \right]^{\frac{1}{2}}$$

$$h_{uv} = \arctan\left(\frac{v^*}{u^*}\right) \qquad (11)$$

The graphical interface of the system provides a heads-up display of three subplots of the $L^*h^*C^*$ color space and allows users to establish color bounds by dragging a graphical rectangle to delineate the desired color range (Figure 2). Color bounds were stable across sub-regions within specimens, and with minor, if any, modifications were applicable across other specimens that had been similarly prepared and stained.

2.2.2.3 Automatic imaging and candidate lymphocyte archiving

The CAM module maps results from the color-filtering operations into exact microscope stage locations for each candidate leukocyte, and directs the robotic scope to systematically image those cells at high magnification. Imaged leukocytes are automatically cropped and transmitted to the client, which systematically segments each cropped image, using a non-parametric clustering algorithm [6, 46], while extracting color, shape, texture and area metrics for each cell's nucleus.

Although lymphocytes may prevail in malignant cases, they normally constitute no more than about 10-15% of the population of leukocytes in peripheral blood. A rejection filter, called the *LymphGate*, was developed to prevent cells other than lymphocytes from entering the database. The rejection filter is based on cell area and roundness, which is computed as

$$roundness = \frac{1}{eccentricity} = \frac{perimeter^2}{4\,\pi\, area} \qquad (12)$$

Cells that fall outside of empirically derived limits are rejected.

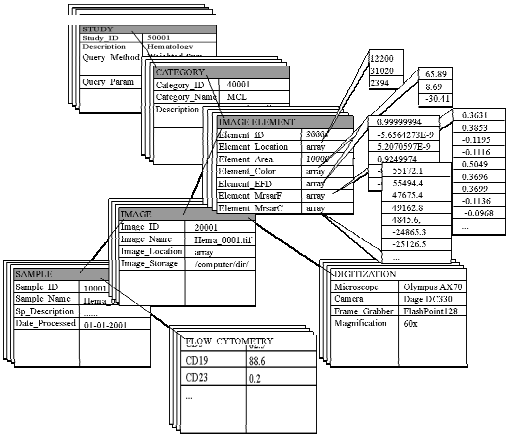

Figure 3. Organization of the ground-truth *PathMiner* database showing the main tables and representative table fields. The major entities are highlighted in gray while auxiliary tables are rendered in black and white. Please see text for more specification.

2.2.3 Image Archiving and Database Management

The *PathMiner* database was designed to archive biomedical images in support of CBIR studies. The database was organized using five fundamental entities: *study*, *category*, *sample*, *image*, and *image-element*. Each *study* is linked to a specific set of auxiliary data, image formats, image elements, image metrics, classifiers and retrieval methods. For each study, there may be several *categories* or classes in which each specific category corresponds to a particular disorder or stage of disease. Each entry in the *sample* entity refers to an actual physical specimen and its correlated clinical data that is stored in auxiliary data tables, e.g., cell surface protein expression profiles and molecular studies. The *image* entity houses information including the image name, a link to the image file, as well as the location of the robotic microscope stage at the time of acquisition. Additional information pertaining to equipment configurations as well as the microscopic magnification is stored in an auxiliary table as shown in Figure 3. The *image-element* refers to region(s) or object(s) of interest within a given image. In the application reported in this paper, the image-elements are the individual lymphocytes. Each image may contain multiple image elements, and each element need not necessarily be classified into the same category, e.g., there may be normal lymphocytes present in malignant lymphoma samples. The constituent entities of the database were designed to be generalizable, modular, and portable thus providing the underlying structure to support a wide range of imaging applications.

The graphical interface for the database provides the means for conducting routine administrative tasks such as creating accounts and assigning privileges to users. The *PathMiner* database interface can be used to populate and manage both local and remote databases. All users have permission to search, index and integrate any databases that reside on their local computer, but such entries do not affect the contents of the ground-truth databases.

2.2.4 Linking lymphocyte images to their molecular characteristics

In addition to archiving the digitized specimen and its corresponding spatial and spectral profiles, the *PathMiner* system incorporates correlated clinical information into the database using customizable auxiliary data tables, thus

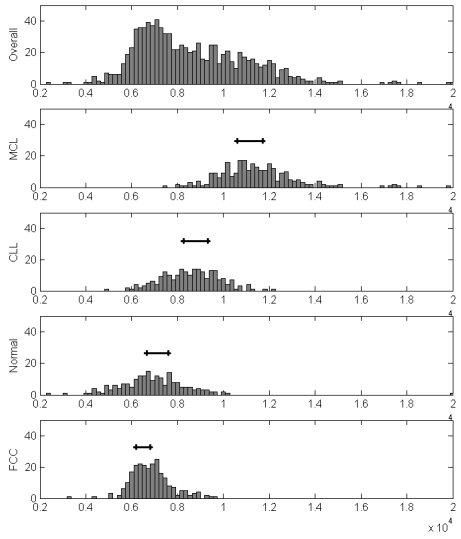

Figure 4. Area histogram of the overall dataset and of each class, showing the size of the estimated bandwidth for each class as bars.

providing a rich resource for investigative research. Specifically designed for this hematology application, a graphical user interface has been developed to enable privileged users to enter and retrieve flow cytometry reports (immunophenotype profiles) to and from the database. This information, along with the corresponding image features, can then be statistically evaluated for patterns and inter-relationships.

Each digital report includes the diagnosis and a brief description of the specimen as well as the gated cell percentage and six possible levels of fluorescence intensity for each protein marker within each cell population. To protect patient privacy, a database index number, which can not be traced back to patient identity, is given to each report so that while patient identity is hidden from users, the correlated protein and/or molecular characteristics of each case can be readily located and retrieved. This interface is actively being utilized by individuals at two institutions (UPenn, UMDNJ) to dynamically expand the ground-truth databases.

A Perl-based program is under development to automatically process flat-ASCII versions of the correlated pathology report by collecting salient fields (diagnosis, immunophenotype, molecular diagnosis, cytochemistry results) while omitting all patient identifiers (Social Security Number, Medical Record Number, Case Accession Number, Address, Phone Number, etc.). The data and identifiers will be developed to meet all HIPAA requirements for sharing data anonymized for research[47].

3 Results

3.1 Classification Optimization

To validate the content based image retrieval methods described in the **Methods** section, two-fold cross validation was performed on a dataset composed of 202 normal, 265 MCL, 223 CLL, and 239 FCC lymphocyte images.

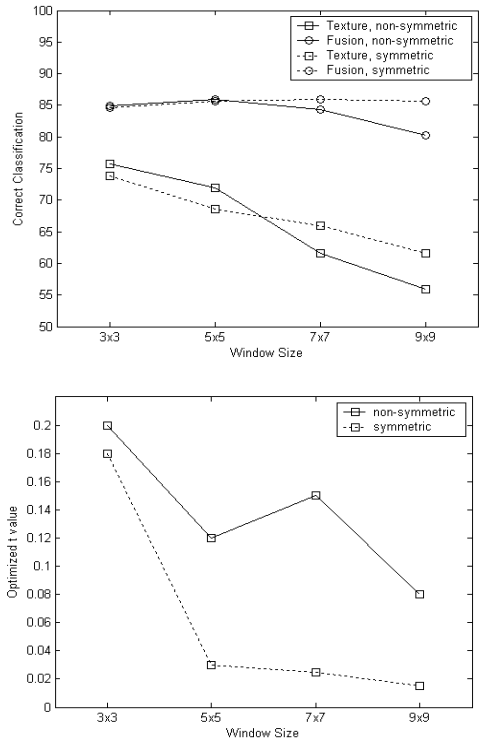

Figure 5. Classification using RSAR texture feature and fusion of RSAR texture with area. a: Classification performance using different window size and neighborhood settings. b: For each neighborhood and window size, t value was optimized to maximize fusion classification performance.

The area feature measurement was the first to be explored in our experiments. Since the area measurement is one dimensional, histograms of the entire mixed test set of cells as well as that of each individual cell class can be visualized as shown in Figure 4. The overall histogram is bimodal because the MCL cells from the test set were larger than the other three classes. The profiles for normal (benign) cells and those for the FCC class exhibited significant overlap, and thus classification using area feature alone would not be optimal. The bandwidth estimation for each class of cells was computed as in (7) above, and was displayed in Figure 4 as a short bar above the mode area. Improvements in the classification performance using area feature alone are plotted in Figure 7.

To test the RSAR method, experiments were first conducted to determine the impact of varying kernel sizes (3x3, 5x5, 7x7 and 9x9), while alternating between two different neighborhood definitions (symmetric and non-symmetric). During the course of these experiments, an optimized bandwidth justifier, t, was chosen to maximize classification performance.

Figure 5a shows that by fusing feature measurements for texture and area, classification performance consistently improved from searching with texture feature alone. These experiments further demonstrated that while smaller window sizes tended to improve the performance when texture metrics were used alone, the rate of correct classification was further improved when a larger kernel window was utilized in conjunction with the fusion of area and texture metrics. A somewhat unexpected result of these experiments was that the symmetric neighborhood provided similar performance to that of the non-symmetric neighborhood in spite of the reduced dimensionality.

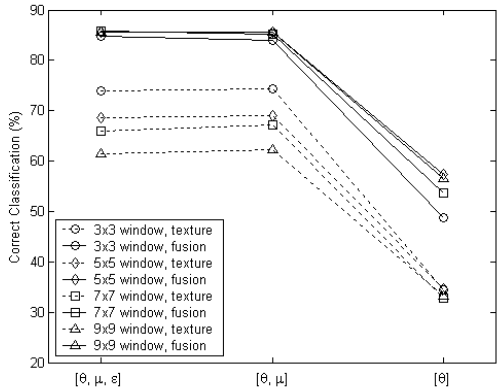

Figure 6. Analysis of classification performance by further reducing dimensions using symmetric RSAR neighborhoods. Dimensionalities of the three situations analyzed are six (the four RSAR parameters θ, the intersect μ, and the residuals ε); five (θ and μ); and four (θ only). Both classification results using the texture feature as well as its fusion with area are shown.

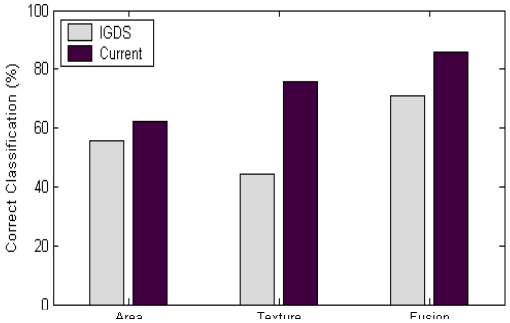

Figure 7. Comparison of classification rate using the original IGDS algorithm and the current system based on two fold cross validation. Classification results based on area feature alone, texture feature alone, and fusion of texture and area are shown. Please see text for an explanation.

The bandwidth justifier, *t*, exhibited a non-dismissible effect on classification performance; the effect is likely to relate to the singularity of the covariance matrix. Although a method for reliably pre-determining its value has not yet been developed, the optimized t values, as shown in Figure 5b, demonstrated that it is possibly related to dimensionality and level of singularity of the texture descriptor.

In an effort to further reduce the dimensionality of the feature space while simultaneously maintaining classification performance, additional experiments were conducted using subdimensions of the symmetric neighborhood in conjunction with the optimized t values, which had been previously computed. From Figure 6 it is evident that while removing the residual dimension does not influence the performance, omission of the intersect dimension has a profound effect. Therefore, the newly developed texture descriptor has a minimum dimensionality of five, rather than the 15 dimensional multi-resolution SAR descriptor used in the previous generation of algorithms of the IGDS system.

Figure 7 compares classification performance in the original and current improved IGDS system based on area, texture feature alone and fusion of both features. It is evident that the new

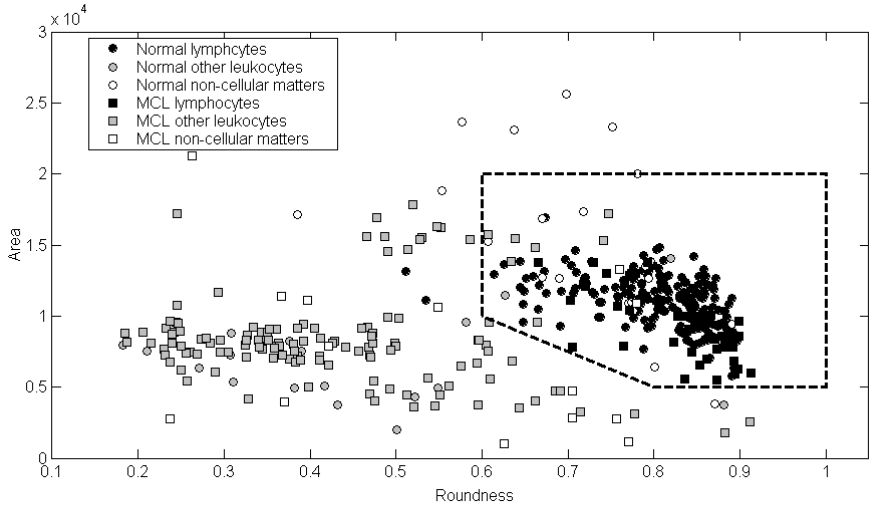

Figure 8. Comparison of the *LymphGate* and human classification result in order to validate feasibility of utilizing the *LymphGate* in discriminating lymphocytes from candidate cell images. Different color and shape of the markers indicate human classification result for each cell image. The *LymphGate* used in this experiment was derived from empirical data and shown in this illustration by the dashed lines. Cell images that fell inside the *LymphGate* were considered candidate lymphocytes and were subject to subsequent steps of IA.

algorithms improved classification significantly in all three categories using the stringent two fold cross validation. Two major factors contributed to the over 30% improvement in texture discrimination. First, the RSAR method was more robust in texture descriptor estimation. Second, by using a carefully selected single-resolution neighborhood instead of the multi-resolution approach we reduced dimensionality of the texture descriptor from fifteen to five. By combining the area and texture feature of the lymphocyte nuclei using probability density based data fusion, the overall correct classification rate for the 929 cells improved by 15%.

3.2 The LymphGate

To test feasibility of the *LymphGate*, the following experiment was set up to compare its discrimination results with human classification. Candidate leukocytes were extracted from a 5x6 frame pilot scan (area of approximately 3mm x 3mm) of peripheral blood

smears from one benign case and one MCL case. Due to high white cell count in the MCL case, only the first 250 out of over 1000 candidate leukocytes were imaged and used for these experiments. Area and roundness features of these cells were plotted in Figure 8, with the *LymphGate*, which was determined from empirical data, outlined by dashed-lines. Cells that fell within the *LymphGate* were considered candidate lymphocytes and were submitted to the subsequent steps of *IA*. Afterwards, a human observer was asked to examine the imaged cells and provide independent classification of each cell image into one of the three groups: lymphocyte; other leukocyte; or non-cellular matter, which includes disintegrated leukocytes, large platelets, and specimen artifacts. The experiment results were plotted in Figure 8. Please note that data from some images from the experiment were not included in Figure 8 for one the following two reasons: the image contains specimen artifacts instead of a cell and hence the segmenter failed to return a valid shape; or the roundness value is larger than one, which implied holes in the segmented shape, a phenomenon specific to neutrophils.

Among the 196 candidate cells extracted from the normal specimen, 40 were classified as lymphocytes by the human observer; all of them, along with 9 cells from other categories, passed the *LymphGate* and were entered into the database. From the MCL specimen, 201 of the 203 lymphocytes classified by human observer passed the *LymphGate*. In the other 15 cells that also passed the *LymphGate*, 13 were classified by human observer as non-cellular matter. They were mostly disintegrated lymphocyte nuclei and thus still bore morphology of lymphocytes. The two lymphocytes that did not pass the *LymphGate* displayed atypical morphology. In summary, the *LymphGate* had an overall sensitivity of 99% and specificity of 85% in our feasibility study.

Cell segmentation and feature extraction processing are completely *unsupervised*. A heads-up display of the output of each step can be observed at the client graphical interface throughout the course of analysis. As an added quality control measure, each indexed candidate lymphocyte is reviewed by a certified hematopathologist prior to its becoming part of the gold-standard database. This is detailed in the next section.

The custom-shaped *LymphGate* was derived from empirical studies rather than generating it directly from known statistical methods for two reasons. First, we were trying to distinguish lymphocytes from non-lymphocytes, which themselves are composed of several different cell groups. As a result, the distribution of the two dimensional data, area and roundness (as shown in Figure 8), does not follow well known linear or quadratic statistical models. Second, the goal was to achieve maximal sensitivity while maintaining a high specificity; this adds yet another level of complexity if the problem was to be adequately addressed with conventional statistical models.

4 Discussion and Future Directions

The chief objective of the *PathMiner* project was to expand upon the progress which had been achieved utilizing the first generation IGDS prototype and establish a large-scale, web-based, resource which could provide reliable decision support for individuals evaluating ambiguous cases in hematopathology. Two approaches were utilized in an attempt to attain this goal. One was to implement the IA system, which facilitates populating the ground-truth database by *unsupervised* scanning and imaging the specimens. The other was to refine the CBIR algorithms to further improve performance and robustness.

The IA subsystem was shown to automatically image, analyze, and archive hematopathology specimens while populating gold-standard databases with resulting image metrics. The IA also provides the means for entering correlated flow reports. In the current stage of development the imaged cells are reviewed by a certified hematopathologist before they become part of the gold-standard databases. Technical highlights of the IA system include the CAM module, color filter, lymphocyte filter, and the image database.

Although our IGDS subsystem was designed as a diagnostic tool, it would have been inappropriate if we had used the IGDS algorithms both for creating the ground-truth database and for testing performance. As a result, a different strategy was used in the *IA* subsystem for selective cell archiving.

The CAM module plays an important part in the IA system by providing the machine equivalent of the "hand-eye coordination" that is normally required to control a robotic system. The CAM module automatically compensates for par-centering and par-focal errors between objective lenses during the pilot scan, thereby providing accurate location of leukocytes for subsequent high-resolution imaging. The color filter and lymphocyte filter allow the IA system to selectively index lymphocytes from peripheral blood specimens. Although it is not impossible to refine the filters so as to eliminate the small number of other leukocytes that yet remains in the candidate cell entries, these cells can easily be picked out by pathologists in the quality control stage of processing and therefore do not affect database integrity. These activities are coordinated by the IA subsystem in order to operate in ***unsupervised*** mode while automatically identifying regions within the specimen, capturing high-resolution images of lymphocytes. The overall advantage to utilizing the CAM module is that it significantly reduces the time and effort required for pathologists to populate and manage the gold-standard database.

The *PathMiner* system provides clinical decision support utilizing a database of cells displaying representative morphology. However, even in confirmed cancer cases, only part of the lymphoid population display typical molecular signatures and morphology for the disease. As a result, the design goal for the IA subsystem at this stage of the project was to present all candidate lymphocytes to pathologists, who would review the cells and include the appropriate ones into the database.

The CBIR algorithms of the IGDS subsystem have undergone major modifications during the course of the *PathMiner* project in order to improve their performance and robustness. Kernel based probability density estimation was used to combine texture and area features or image retrieval. Compared to the MRSAR method used in previous generation of algorithms used in the IGDS, the newly developed RSAR approach was proven to be more robust in describing regional texture and computationally more efficient.

As demonstrated in our study, the 5x5 symmetrical neighborhood configuration was the most effective approach for capturing distinctive textural features of each class and complementing the area classification. By further removing the residual dimension, dimensionality of texture descriptor was reduced from 15-dimensional to the final five-dimensional, which should be the most important contributing factor in improving the overall robustness of the system. The classification performance based on two fold cross validation using 929 cells increased by more than 15%.

The *PathMiner* system reported in this paper classifies individual lymphocytes according to their morphologic signatures. The next phase of development for the *PathMiner* project is to expand the number and type of malignancies under study while continuing to improve the underlying machine vision and CBIR algorithms. Future plans for the *PathMiner* system include generating a diagnosis based upon composite or aggregate analysis of lymphocytes across the entire specimen.

Acknowledgements

This work was supported in part by NIH contract 5 RO1 LM007455-02 from the National Library of Medicine.

References:

[1] J. Garcia-Conde and F. Cabanillas, Mantle cell lymphoma: a lymphoproliferative disorder associated with aberrant function of the cell cycle, Leukemia 10 Suppl 2 (1996) s78-83.
[2] G. Vadlamudi, K. A. Lionetti, S. Greenberg and K. Mehta, Leukemic phase of mantle cell lymphoma two case reports and review of the literature, Arch Pathol Lab Med 120 (1996) 35-40.
[3] J. K. Chan, P. M. Banks, M. L. Cleary, G. Delsol, C. De Wolf-Peeters, B. Falini, K. C. Gatter, T. M. Grogan, N. L. Harris and P. G. Isaacson, A revised European-American classification of lymphoid neoplasms proposed by the International Lymphoma Study Group. A summary version, Am. J. Clin. Pathol. 103 (1995) 543-60.
[4] Y. Yatabe, S. Nakamura, M. Seto, H. Kuroda, Y. Kagami, R. Suzuki, M. Ogura, M. Kojima, T. Koshikawa, R. Ueda and T. Suchi, Clinicopathologic study of PRAD1/cyclin D1 overexpressing lymphoma with special reference to mantle cell lymphoma. A distinct molecular pathologic entity, Am J Surg Pathol 20 (1996) 1110-22.
[5] C. Rozman and E. Montserrat, Chronic lymphocytic leukemia, N Engl J Med 333 (1995) 1052-7.
[6] D. J. Foran, D. Comaniciu, P. Meer and L. A. Goodell, Computer-assisted discrimination among malignant lymphomas and leukemia using immunophenotyping, intelligent image repositories, and telemicroscopy, IEEE Trans Inf Technol Biomed 4 (2000) 265-73.
[7] M. N. Kilo and D. M. Dorfman, The utility of flow cytometric immunophenotypic analysis in the distinction of small lymphocytic lymphoma/chronic lymphocytic leukemia from mantle cell lymphoma, Am J Clin Pathol 105 (1996) 451-7.
[8] C. H. Geisler, J. K. Larsen, N. E. Hansen, M. M. Hansen, B. E. Christensen, B. Lund, H. Nielsen, T. Plesner, K. Thorling, E. Andersen and et al., Prognostic importance of flow cytometric immunophenotyping of 540 consecutive patients with B-cell chronic lymphocytic leukemia, Blood 78 (1991) 1795-802.
[9] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick and C. Sultan, Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group, Br J Haematol 33 (1976) 451-8.
[10] D. R. Head, R. A. Savage, L. Cerezo, C. M. Craven, J. N. Bickers, R. Hartsock, T. A. Hosty, J. H. Saiki, H. E. Wilson, F. S. Morrison and et al., Reproducibility of the French-American-British classification of acute leukemia: the Southwest Oncology Group Experience, Am J Hematol 18 (1985) 47-57.
[11] H. Harms, H. M. Aus, M. Haucke and U. Gunzer, Segmentation of stained blood cell images measured at high scanning density with high magnification and high numerical aperture optics, Cytometry 7 (1986) 522-31.
[12] I. Baumann, R. Nenninger, H. Harms, H. Zwierzina, K. Wilms, A. C. Feller, V. Ter Meulen and H. K. Muller-Hermelink, Image analysis detects lineage-specific morphologic markers in leukemic blast cells, Am J Clin Pathol 105 (1996) 23-30.
[13] S. Kneitz, G. Ott, R. Albert, T. Schindewolf, H. K. Muller-Hermelink and H. Harms, Differentiation of low grade non-Hodgkin's lymphoma by digital image processing, Anal Quant Cytol Histol 18 (1996) 121-8.

[14] O. Debeir, C. Decaestecker, J. L. Pasteels, I. Salmon, R. Kiss and P. Van Ham, Computer-assisted analysis of epiluminescence microscopy images of pigmented skin lesions, Cytometry 37 (1999) 255-66.
[15] M. Cenci, C. Nagar and A. Vecchione, PAPNET-assisted primary screening of conventional cervical smears, Anticancer Res 20 (2000) 3887-9.
[16] 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 24 (2001) 426-34.
[17] P. H. Bartels, R. M. Montironi, D. Bostwick, J. Marshall, D. Thompson, H. G. Bartels and D. Kelley, Karyometry of secretory cell nuclei in high-grade PIN lesions, Prostate 48 (2001) 144-55.
[18] J. W. Smith, Jr., J. R. Svirbely, C. A. Evans, P. Strohm, J. R. Josephson and M. Tanner, RED: a red-cell antibody identification expert module, J Med Syst 9 (1985) 121-38.
[19] D. R. Thursh, F. Mabry and A. H. Levy, Computers and videodiscs in pathology education: ECLIPS as an example of one approach, Hum Pathol 17 (1986) 216-8.
[20] B. N. Nathwani, K. Clarke, T. Lincoln, C. Berard, C. Taylor, K. C. Ng, R. Patil, M. C. Pike and S. P. Azen, Evaluation of an expert system on lymph node pathology, Hum Pathol 28 (1997) 1097-110.
[21] M. Das, E. Riseman and B. Draper, FOCUS: Searching for multi-colored objects in a diverse imag database., IEEE Conf. on Comp. Vis. and Pattern Recognition, San Juan, Puerto Rico, (1997), 756-761.
[22] M. Flickner and e. al., Query by image and video content: The QBIC system, Computer 9 (1995) 23-31.
[23] W. Ma and B. Manjunath, Texture-based pattern retrieval from image databases., Multimedia Tools and Applications 1 (1996) 35-51.
[24] M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra and T. Huang, Supporting similarity quries in MARS, ACM Multimedia, Seattle, Washington, (1997), 403-4131.
[25] S. Antani, R. Kasturi and R. Jain, Pattern recognition methods in image and video databases: past, present, and future, in Advances in Pattern Recognition, Lecture notes in comp science, eds. A. Amin, pp. 31-53 (Springer-Verlag, London, UK, 1998).
[26] C. Faloutsos, R. Barber, M. Flickner, W. Hafner, W. Niblack and e. al., Efficient and effective querying by image content., J. of Intelligent Info. Sys:Integrated Artificial Intelligentce and Database Technologies. 3 (1994) 231-262.
[27] A. Pentland, R. Picard and S. Sclaroff, Photobook: content based manipulation of image databases, Int. J. Comp. Vis. 18 (1996) 233-254.
[28] J. Wang, G. Wiederhold, O. Firschein and S. Wei, Wavelet-based image indexing techniques with partial sketch retrieval capability, Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries, ADL, Washington, DC, (1997) 13-24.
[29] C. Carson, M. Thomas, S. Belongie, J. Hellerstein and J. Malik, Blobworld: a system for region-based image indexing and retrieval, Third International Conference on Visual Information Systems, Amsterdam, the Netherlands. (1999).
[30] J. Li, J. Wang and G. Widerhold, IRG: Integrated Region Matching for image retrieval, Proc. of ACM Multimedia, Los Angeles, (2000),
[31] F. Schnorrenberg, C. Pattichis, C. Schizas and K. Kyriacou, Content-based retrieval of breast cancer biopsy slides., Technology & Health Care 8 (2000) 291-7.
[32] C. Le Bozec, E. Zapletal, M. Jaulent, D. Heudes and P. Degoulet, Towards content based image retrieval in a HIS-integrated PACS., Proceedings / AMIA Annual Symposium (2000) 477-481.
[33] M. Jaulent, A. Bennani, C. Le Bozec, E. Zapletal and P. Degoulet, A customizable similarity measure betwen histological cases, Proceedings / AMIA Annual Symposium (2002) 350-354.
[34] A. Wetzel, R. Crowley, S. Kim, R. Dawson, L. Zheng, Y. Joo, Y. Yagi, J. Gilbertson, C. Gadd, D. Deerfield and M. Becich, Evaluation of prostate tumor grades by content based image retrieval, Proceedings of Spie -- the International Society for Optical Engineering 3584 (1999) 244-245.
[35] J. Wang, Pathfinder: multiresolution region-based searching of pathology images using IRM., Proceedings / AMIA Annual Symposium (2000) 883-887.
[36] M. E. Mattie, L. Staib, E. Stratmann, H. D. Tagare, J. Duncan and P. L. Miller, PathMaster: content-based cell image retrieval using automated feature extraction, J Am Med Inform Assoc 7 (2000) 404-15.
[37] W. Press, S. Teukolesky, W. Vetterling and B. Flannery, Numerical recipes in C: the art of scientific computing, Chap. 10 (Cambridge University Press, New York, 1992).
[38] R. Duda, P. Hart and D. Stork, Pattern Classification, second edition, (2001)
[39] D. Comaniciu, P. Meer and D. J. Foran, Image-guided decision support system for pathology, Machine Vision and Applications 11 (1999) 213-224.
[40] J. Mao and A. Jain, Texture classification and segmentation using multiresolution simultaneous autoregressive models, Pattern Recognition 25 (1992) 173-188.
[41] R. Bellman, Adaptive Control Processes: A Guided Tour., in eds. pp. (Princeton University Press, 1961).
[42] M. Koeppen, The Curse of Dimensionality, 5th Online World Conference on Soft Computing in Industrial Applications (WSC5), held on the Internet, (2000),
[43] H. L. Chen, I. Shimshomi and P. Meer, Model based object recognition by robust information fusion., 17th International Conference on Pattern Recognition, Cambridge, U.K., (2004),
[44] T. Cover and J. Thomas, Elements of Information Theory. (John Wiley, New York, 1991).
[45] G. Wyszecki and W. Stiles, Color Science: Concepts and Methods, Quantitative Data and Formulai, 2nd edition (John Wiley & Sons, Inc., 1982).
[46] D. Comaniciu and P. Meer, Cell image segmentation in diagnostic pathology, in Advanced Algorithmic Approaches to Medial Image Segmentation: State-Of-The-Art Applications in Cardiology, Neurology, Mammography and Pathology, eds. J. Suri, S. Singh and K. Setarehdam, pp. 541-558 (Springer, 2001).
[47] J. J. Berman, Concept-match medical data scrubbing: How pathology text can be used in research, Arch Pathol Lab Med 127 (2003) 680-686.