Scientific Research and Essays Vol. 6(2), pp. 247-255, 18 January, 2011
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE10.482
ISSN 1992-2248 ©2011 Academic Journals

Full Length Research Paper

Dual modality search and retrieval technique analysis for leukemic information system

S. Rajendran[1], H. Arof[1], N. Mokhtar[1]*, M. Mubin[1], S. Yegappan[2] and F. Ibrahim[3]

[1]Department of Electrical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia.
[2]Hospital Ampang, Selangor, Malaysia.
[3]Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia.

Accepted 11 November, 2010

Medical Information System (MIS) deals with standardized method of collection, storage, retrieval and evaluation of patient data. Computational Intelligence (CI) technique has many abilities in data processing and structuring, pattern matching, forming knowledge base, reasoning and decision making. MIS for leukemia cells was developed at University of Malaya for the Hematology Department of Hospital Ampang, Malaysia. CI techniques were used for syntactical and contextual image retrieval of leukemic images. A logical approach was used to systematically develop content based image retrieval system that supports decision making in clinical hematopathology. The developed system is based on open source. Pages were designed and developed using Java Server Pages (JSP) and Java with MySQL as the database for the domain and image repository. Several Java-based tools were used for image processing, neural network based pattern classification and recognition, and for the hybrid system. The classification rates for individual white blood cells varied between 97% for training data and 95% for testing data depending on the individual cell type- monocyte, lymphocyte, eosinophil, basophil and neutrophil. This result is cross-validated by an expert hematologist.

Key words: Medical information system (MIS), leukemia image retrieval, image processing, feature classification, pattern recognition, hematology.

INTRODUCTION

The exponential growth of intense research and development in various fields (particularly computer vision) has resulted in rapid technological advancements in the field of medicine. Medical images play a vital role in the medical diagnoses, research and teaching. Digitization of the medical images and further analysis of the images provides a scope for the enhancement of the depth of the diagnosis in terms of accuracy, measurability, sensitivity and future reference with quality assurance.

The application of various techniques of computer vision in the medical field has helped doctors, researchers, specialists and patients in many ways. Medical images can be stored and retrieved in digital format. They can be reconstructed in 3 dimensional (3D) form for computer tomography of human body and can be used as models for fabricating the maxillofacial parts for the oral cancer patients and many more. There are imaging studies on medical images like X-ray, Magnetic Resonance Imaging (MRI), radionuclide scanning and ultrasound to determine if cancer (lymphoma and leukemia) has invaded other organs in the body. Leukemia is the subject of interest for our research. A brief overview about Leukemia and a review of similar relevant researches from the domain perspective and their technicalities are discussed below.

Leukemia is a group of serious blood diseases affecting white blood cells or leukocytes in the bone marrow. Leukemia leads to an overproduction of abnormal or immature white blood cells that reduces the ability in fight against infection (Rozenberg, 2003). There have been numerous works done in the field of hematology and

*Corresponding author. E- mail: norrimamokhtar@um.edu.my. Tel: +6012-2285060, +603-79674599.

some of them are described as follows:

(i) Yan et al. (2006) proposed novel hybrid merging algorithm that could combine the probability density function scores on fragments to classify circular and non-circular cells. They developed a system that has a traditional watershed algorithm for the cell nuclei segmentation and phase identification for automated analysis that can deal with large volume of imaging from a time-lapse optical microscopy.
(ii) Demir and Yener (2005) presented a systematic survey on the computational steps that are involved for automatic classification and analysis of cancer diagnosis system based on histopathology images. Feature selection procedures and feature extraction techniques with respect to the cancer cells were extensively surveyed and discussed. Intensity values of the pixels and their spatial interdependency were performed to extract these features.
(iii) Walker (1997, 2007) developed an automated texture analysis system based on the self-adaptive texture analysis technique for his cytology images. Bhattacharya discrimination (Bhattacharya, 1943) measure was used for classification.
(iv) Olivier and Vega (2000) discussed about morphological features of the lymph node that were used for performing a feature space search and similarity measures for their expert system.
(v) Beksaç et al. (1997) were the pioneers in developing diagnostic aid systems for microscopic histological cell images. The step-by-step approach for developing the system was quite impressive; however the density and texture specifications were not considered during the feature extraction.

In summary, these medical images have gained potential importance in the decision making, diagnosis and prognosis. There is a perpetual growth in application of medical multimedia data. Medical information systems and diagnostic aid systems are becoming a necessity more than sophistication. For example, our research hospital was lacking from a proper storage and retrieval systems related to leukemia. Traditional microscopic analysis has to be carried out repeatedly whenever there is a need for more information.

This research is an initiative to address the above issue. We designed and developed a medical information system of leukemia for our research hospital. This system is expected to be used as a storage and decision support by the hematologists for diagnosis, prognosis and decision making. Our aim is to develop a storage system for the patient details, to search and retrieve in dual modality (keyword and image), classify individual leukocytes for the knowledge base, to perform knowledge engineering for the base, and thereby develop a diagnostic aid system.

One of the researches that were developed with similar ideas and techniques was Hengen et al. (2002). They worked with bone marrow while we chose to analyze using peripheral blood smear since the individual cells are not clustered. The benefits of our system are as follows:

(i) Storage, retrieval and decision support systems reduce the time and effort taken by the hematologists in their routine diagnosis.
(ii) Storage of digitized microscopic images will last longer and can be used for future references.
(iii) Digitized images enable doctors to observe patient's prognosis for clinical treatment.

System can be a virtual guide for inexperienced hematologists, assistants and other researchers in hematology laboratory.

MATERIALS AND METHODS

Our MIS for Leukemia consists of two main modules as follow:

(i) Database management system for patients, users, images (patients' diseased cells) and the features of the images.
(ii) Search and retrieval systems based on text and image.

In formulating the database system we have utilized pre-processing, image processing, computational intelligence and search techniques (Bishop, 1996; Comaniciu et al., 1998; Comaniciu et al., 2001; Demir et al., 2005; Giuseppe et al., 2005; Wang et al., 2006). The system was developed with core Java, Java Server Pages (JSP), Javascript, MySQL database, imagej-Java based, image visualisation software (Rasband, 2007) and Weka–Data mining software [Ian et al., 2005].

System design

The structure of the MIS is divided into three layers based on its components and functionalities as shown in Figure 1. The three layers are the physical storage, content abstraction and the front-end interface layers. The physical storage layer itself has a few database systems catering to store each type of data.

Patient database system contains all the details about the patient like serial code, name, age, sex, race, case history, diagnosis and other required details. Image database system consists of serial code of the patient, images and thumbnail of the image.

Image processing, feature extraction and feature classification are performed to classify the features into a set of known feature vectors.

Feature vector database system consists of the feature vectors represented both in numerical and structural representation. Feature vector includes the shape of the nucleus, cytoplasm, presence of nucleolus or nucleoli, perimeter and area of cytoplasm and nucleus, nucleus/cytoplasm ratio and numerical calculations like mean, variance, standard deviation for nucleus and cytoplasm.

Knowledge base consists of storage of all the individual white blood cells with their name and characteristics as their semantic annotation. Rule based reasoning is incorporated to perform action on the priori knowledge to infer the result.

The content abstraction layer is the middle layer that acts as the backbone of the system and performs all the commands sent to it and then sends back the result to the interface layer.

The interface layer is the front-end layer that communicates with the user. Interface layer provides a web page or any environment to browse, search or upload in the system. The layer also displays the retrieved result, which is relevant to the query.

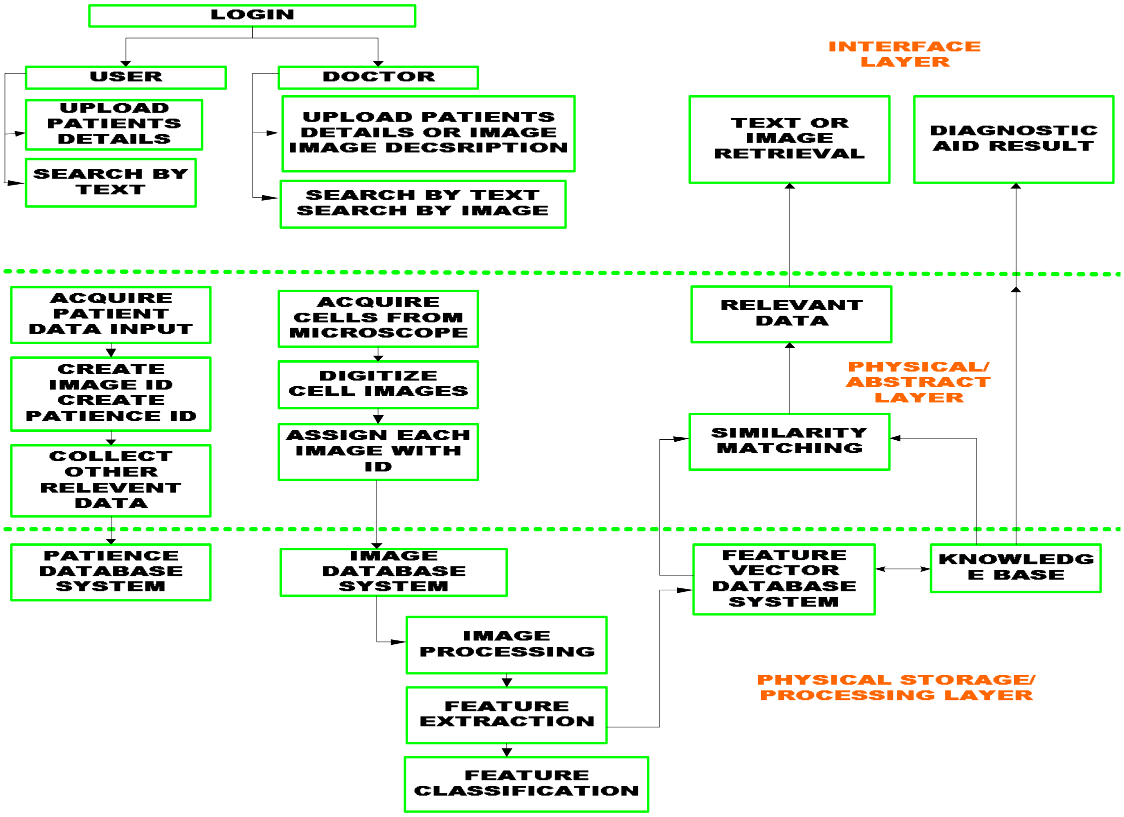


Figure 1. Structure of the medical information system.

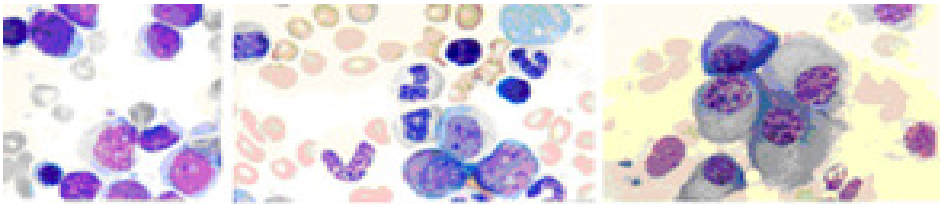


Figure 2. Sample of acquired leukemia images.

Data modeling and management

In this system, database consists of patient table, image table and feature table. Patient identification (ID) is a non-zero primary key set to the patient table. Image ID is the foreign key for the patient table and primary key for the image table. Both Patient ID and Image ID refer to the unique serial number assigned to a record. Feature table parameters are dimensionally modeled so that the feature vectors can be sliced from the image for pattern classification and recognition.

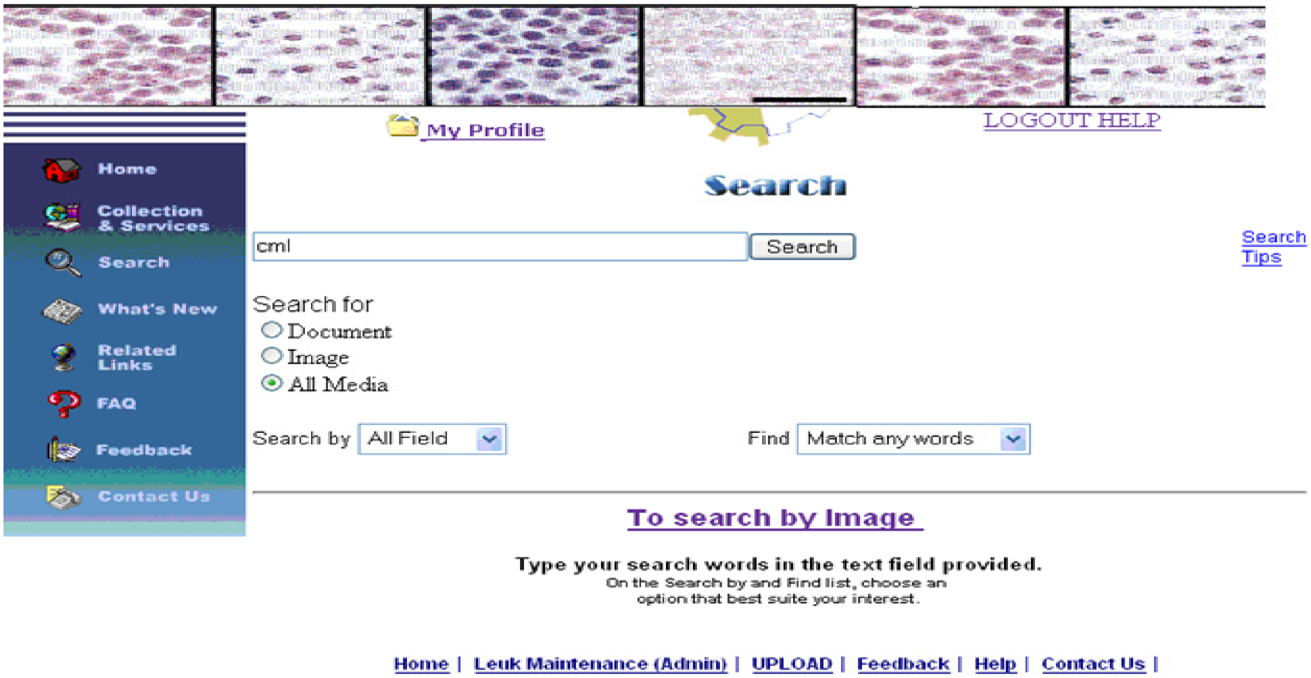
Image acquisition and standardisation

Analog images have been acquired from Olympus BX-50 microscope with a magnification of 3.3×100. Olympus SC35 analog camera with frame grabber PA1-10A was used to capture the micro images. These analog images were digitized with a HP -1510 digitizer with a resolution of 400 dpi. The obtained digitized images were verified with the hematologist for the microscopic details of the features and contrast of the individual cell of the image. Figure 2 shows the digitized images that were obtained from Ampang Hospital.

SEARCH AND RETRIEVAL SYSTEMS

Search techniques

Two search techniques were used in the patient database management system. A basic string or exhaustive search using Brute-force algorithm (http://en.wikipedia.org/wiki/Brute-force_search] was used for text–based search for retrieving exact match of patient records. Heuristic-based beam search (http://en.wikipedia.org/wiki/Beam_search) was used for image-based search to retrieve best-first similar leukemia images from the database. Figure 3 shows the text and image based search interface.



cml

Search for
○ Document
○ Image
● All Media

Search by All Field

Find Match any words

Type your search words in the text field provided.
On the Search by and Find list, choose an option that best suite your interest.

Figure 3. Text and image based search interface

Indexing technique

Indexing is carried out at word and record levels for text retrieval. While for image retrieval, indexing is based on secondary keys with one or more attributes. Data and queries are considered as vectors of multidimensional feature space. Feature extraction computes feature vectors and indexing organizes the feature space appropriately so that it can answer queries on any attribute. Quad-tree image representation was applied to represent recursive subdivisions of an image. Advantage of Quad tree image representation is that they are capable of handling disjoint structures like cytoplasm, nucleus, and nucleolus in a cell.

Retrieval techniques

The retrieval technique was used based on region matching along with semantic annotation. Region matching was preferred over pixel matching as region matching can maintain the invariance of rotation, translation, scaling and intensity microscopic images. Similarity measure can be obtained from region matching while semantic annotation can classify the feature vectors into specific categories that are also used for feature indexing. Low level features like color, texture, shape and size are extracted and classified as individual feature vectors and then the region to region matching is performed based on the feature vectors. The weighted sum of region to region matching gives the distance between images.

Retrieval by metadata

Exact match by metadata was performed to retrieve the patient details along with thumbnail images. String matching by brute-force algorithm is illustrated as follows:

JSP code snippet

```
Void BF (char x, int m, char y, int n)   int i, j;

 /* Searching */
 for (j = 0; j <= n - m; ++j)
 for (i = 0; i < m && x[i] == y[i + j]; ++i);
 if (i >= m)
 print(j);
```

Retrieval by similarity

Image retrieval involves image processing, feature classification and organization, indexing feature vector structure and pattern matching techniques.

Image processing techniques

Gaussian filters were used to remove irrelevant objects in the background like the red blood cells. Java image processing toolkit imagej (Rasband, 2007) was used in this study and Figure 5 depicted the process. K-means clustering was used for segmenting nucleus and cytoplasm. After segmentation, the next step is to extract the features based on their boundaries. Features like shape of the nucleus, cytoplasm and geometrical properties were extracted. These features are used as a vector for feature indexing.

Followed by feature extraction, particle analysis was performed. The size range of the particles is given in the input and those which are above or below the range are ignored and the circularity of the particle is also filtered using the range. The circularity can be calculated using the formula 4pi (area/perimeter^2) and a value of 1 indicates a perfect circle.

Feature vector analysis was performed after the particle analysis. Finding quantitative or structural descriptive features in blood cell image is the one of the most important issues for analyzing an image. This analytical process includes quantification, extraction and selection of features of the cell images. Every characteristic

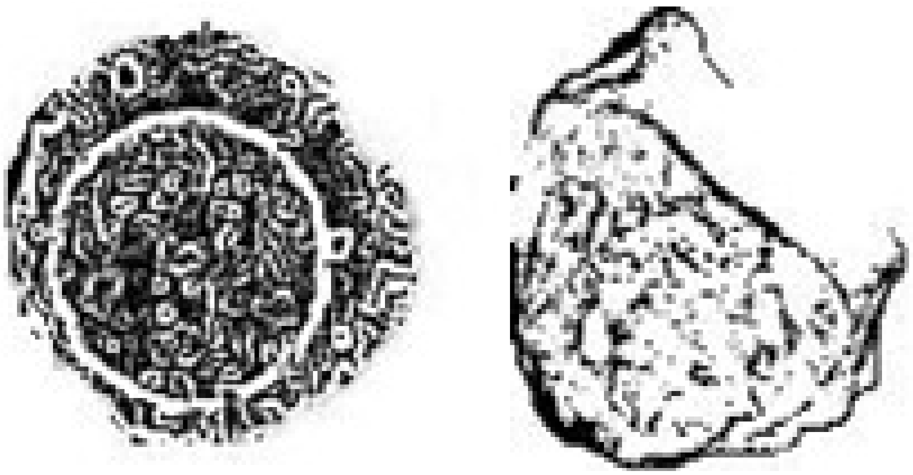


Figure 4. Myeloid and lymphoid cells image source: Hospital Ampang (Courtesy: image J).

and nature of image features contributes in forming a conclusion about the type and severity of leukemia. Therefore analyzing the image based on feature vectors is crucial. To obtain the feature vectors, features were analyzed based on low-level features like color, geometry and texture and high level features like semantic annotations.

Feature classification

Feature classification is a statistical procedure in which individual items are grouped together based on their inherent quantitative information. Lazy classifier with K-Nearest Neighbor algorithm was performed for the pattern recognition of the image regions. The K-Nearest Neighbor algorithm classified the objects based on observable features like shape that are closest to the training samples of the space. The algorithm selects a set which contains the K-Nearest Neighbors and assigns the class label to the new data point based upon the most numerous classes within the set.

Feature organization

A feature vector is then created from various computed features and organized into a data structure for efficient retrieval. Features are stored as numerical and structural parameters in the feature table of the database. The feature organization strategy is strongly dependent on the feature vector used, the query types supported and the image annotation or semantics. We are at a stage where we have some of these requirements identified. Although the research works on getting the most efficient strategy for the feature vector organization, it also has flat file structure and text search for retrieval purposes.

Correlation of feature vectors with diagnosis of leukemia blood cells

Lymphoid series and myeloid series can be differentiated by the shape of the nucleus primarily. Myeloid has a circular nucleus whereas lymphoid does not have a regular shape as observed in Figure 4.

The nucleus of myeloid is denser compared to lymphoid. Acute and chronic cases can be well differentiated by the size of the individual cells present in the field.

Pattern matching by chamfer distance transforms

The regions of the images are matched based on the chamfer distance transform. Chamfer algorithm first takes out the edges and filters out the noisy and weak edges, then with the edges distance transform is obtained between the query image and the other relevant images. Using the distance transform, matching between two images is calculated. The least pixel distance gives the best matching result and that means that image is the closest to the query image.

RESULTS

The results are analyzed and presented based on three different perspectives namely the technical, medical and utility perspectives. Technical perspective presents the performance and evaluation of the system based on its technicalities used.

Medical perspective presents the opinion of the hematologist about the developed system, its advantages and problems. Utility perspective presents the view about the interface for navigating through the system, information access and availability.

Technical perspective

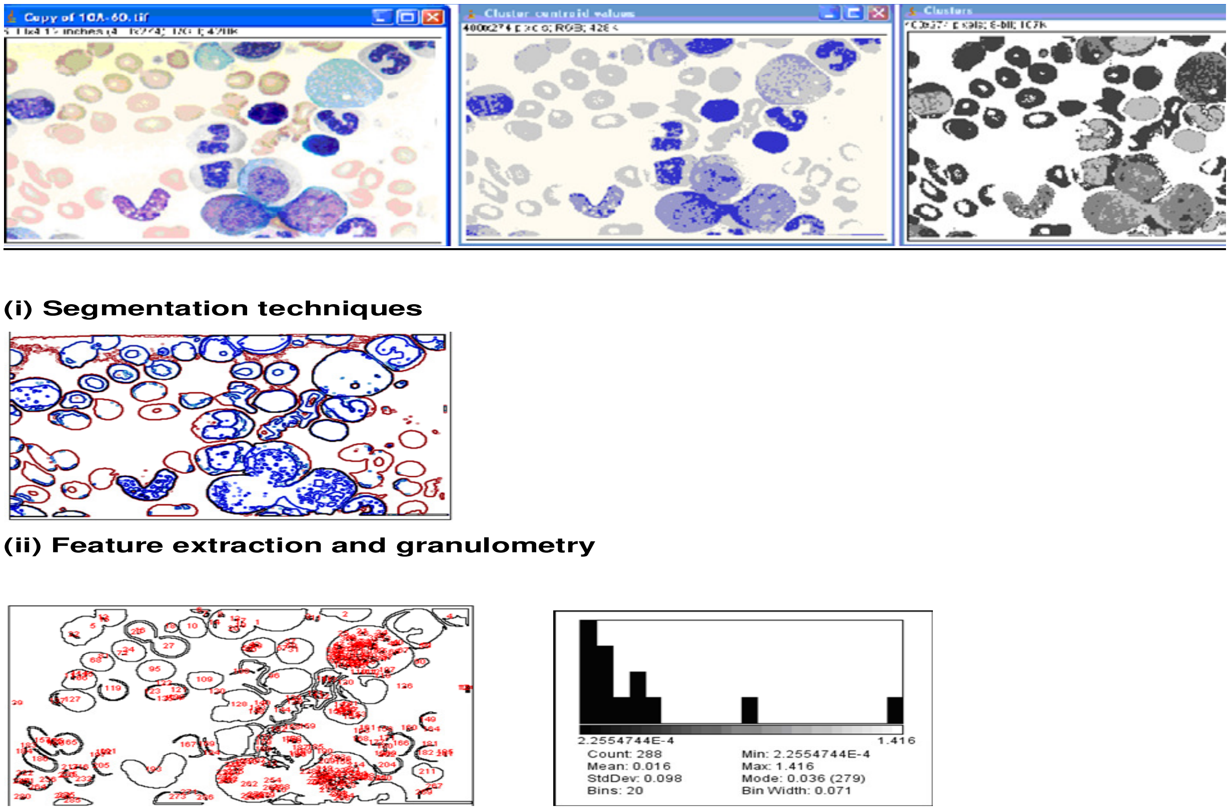
The technical perspective is obtained from the image processing, clustering, classification analysis and retrieval results.

Processing analysis

The image processing results are analyzed for segmentation and feature extraction techniques.

Segmentation

Cytoplasm and nucleus was segmented accurately using



(i) Segmentation techniques

(ii) Feature extraction and granulometry

(iii) Particle analysis and particle size distribution

Figure 5. Image processing techniques.

edge based segmentation algorithm and the over-segmentations were filtered through the maximum filter in order to filter out the red blood cells if present in the background. The processing time was less than 1.2 s.

Feature extraction

Features were extracted based on shape, size and texture. Feature vectors were formed for:
Shape (circularity), Size (cell area, cell perimeter) and Texture (angular second moment, contrast, correlation and entropy).

Feature vector organization proved to be useful for retrieving relevant images but there was a dimensionality curse as computing all these features took a long time and so quad tree indexing structure was introduced to reduce the retrieval time. Image retrieval took about 2 to 3 s for retrieving 5 similar images. This can be further improved if pruning technique is incorporated to the beam search algorithm.

Attribute selection

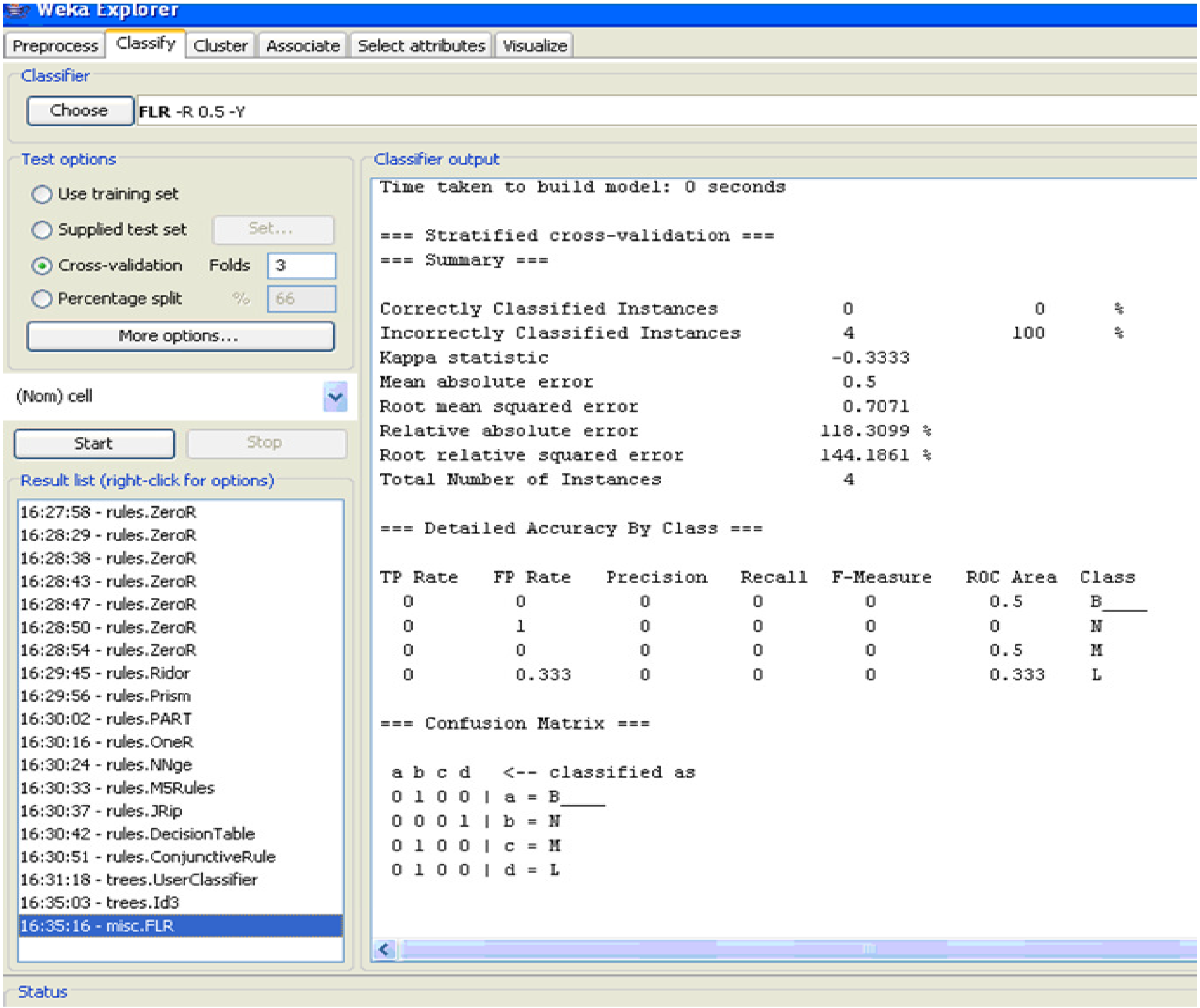
The selection of attributes along with the weight for each attribute was evaluated using the “CFssubset” evaluating algorithm using a 3-fold cross validation. Six attributes were selected for analysis – cell, area, angular second moment, contrast, correlation and entropy.

Cell clustering analysis

The image processing and attribute selection was followed by the image analysis based on clustering and classification. Density based cluster was used for evaluation based on the training data to cluster based on their cell attributes using Weka. Clustering was performed for two iterations. Figure 5 displayed the segmentation techniques, feature extraction and granulometry, and particle analysis and particle size distribution.

Cell classification

The confusion matrix of Figure 6 shows 1 on each column for every row and this classifies the individual cells. The interpretation of the result where a, b, c, d are assigned as B (basophil), N (neutrophil), M (myelocyte)



FLR -R 0.5 -Y

○ Use training set
○ Supplied test set
● Cross-validation Folds 3
○ Percentage split % 66

(Nom) cell

16:27:58 - rules.ZeroR
16:28:29 - rules.ZeroR
16:28:38 - rules.ZeroR
16:28:43 - rules.ZeroR
16:28:47 - rules.ZeroR
16:28:50 - rules.ZeroR
16:28:54 - rules.ZeroR
16:29:45 - rules.Ridor
16:29:56 - rules.Prism
16:30:02 - rules.PART
16:30:16 - rules.OneR
16:30:24 - rules.NNge
16:30:33 - rules.M5Rules
16:30:37 - rules.JRip
16:30:42 - rules.DecisionTable
16:30:51 - rules.ConjunctiveRule
16:31:18 - trees.UserClassifier
16:35:03 - trees.Id3
16:35:16 - misc.FLR

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           0               0      %
Incorrectly Classified Instances         4             100      %
Kappa statistic                         -0.3333
Mean absolute error                      0.5
Root mean squared error                  0.7071
Relative absolute error                118.3099 %
Root relative squared error            144.1861 %
Total Number of Instances                4

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0         0          0          0         0         0.5      B____
  0         1          0          0         0         0        N
  0         0          0          0         0         0.5      M
  0         0.333      0          0         0         0.333    L

=== Confusion Matrix ===

 a b c d   <-- classified as
 0 1 0 0 | a = B____
 0 0 0 1 | b = N
 0 1 0 0 | c = M
 0 1 0 0 | d = L
```

Figure 6. Classification of individual cells. Courtesy :Weka.

and L (lymphocyte), respectively.

The classification rates were 95% for the training data. The classification rate can further be improved to 98% if the color parameter is also taken under consideration. The classification rates obtained for testing data were 95% for a 3 fold cross validation.

RETRIEVAL ANALYSIS

Text retrieval

The performance of text retrieval was evaluated in two ways:

(i) Retrieval time taken.
(ii) Exactness of the match

The time taken for retrieving the document was consistently less than 0.3 s and ten documents were placed in a page with the text of the document or patient case report on the left side and the multimedia document like image on the right side. Exactness of the match was remarkably good, as every time during testing, the first page or the first ten documents were the most relevant.

Image retrieval

The performance of image retrieval can be evaluated in terms of precision. The diagnosis of leukemic type can be evaluated based on specificity and sensitivity. Precision of the retrieval was on an average 0.7693 (76.93%). Every 13 images retrieved only 9 of them were relevant.

Recall was 75% as it retrieved 6 relevant images out of 8 relevant images available in the database. The other 2 images though belonged to the same type of leukemia were not similar in pattern. Realizing the balance between specificity and sensitivity with respect to the hematologist's acceptance level is beyond the scope of this project.

From medical perspective, the differentiation of one image to other may vary in many cases based on the sensitivity and specificity of that particular image. This accounts for the calibration variation between technical and medical experts. However, since this system has to help the clinicians, it has to be further improved in accordance with their perspective. The pattern search algorithms have to be custom-made to improve the overall performance. Pruning techniques can further enhance for the reduction in retrieval time and complexity.

MEDICAL PERSPECTIVE

The system is user friendly for the hematologist to navigate, browse and retrieve the old patient records. According to the medical expert and hematologist at Hospital Ampang, the system is able to differentiate the acute and chronic cases very well. The classification of individual white blood cell is good. The overall MIS for the similarity based image retrieval, retrieval time and relevance is accurate. The main advantage of the system is that the medical doctors can compare patients' old leukemia images to the latest images in determining the progress of the disease during treatment.

UTILITY PERSPECTIVE

Utility perspective is about performance of the developed system based on the interface design, availability of information and access to that data. There are three levels of security access to the system; administrators, medical experts and other users. Administrators control the management of the data storage, retrieval, authentication, tracking and any other issues related to the system. The medical experts have access to the patient's medical images for uploading, browsing, downloading, comparing and analyzing. The computer users are only granted the permission to upload the patient records, browse and search the records based on text from the patient database.

DISCUSSION

The system needed an efficient combination of content based and text based retrieval for a database of image and flat files. The data was modelled according to the domain, context and feature space relationship was drawn between these spaces.

Only three levels of features (cytoplasm, nucleus and nucleolus) were selected to reduce the complexity. One more level–vacuoles, of further improvement could help in producing few distinct results as the presence of vacuoles could determine specific types.

Texture and geometric features helped in individual white blood cell classification. However, an addition of color feature could improve the classification rate.

Balancing the specificity and sensitivity, classifying the subclasses were difficult. Heuristic rules have to be introduced to address these problems. The real challenge is to improve the specificity, sensitivity and to set a balance between simplicity and complexity.

The indexing structure for text retrieval performed well, however indexing structure formed for the feature sets performance need to be improved for medical domain. The solution is to consider the colour feature sets as a feature vector for indexing and classification.

Conclusion

Initiative taken for developing leukemia MIS was achieved. The system was evaluated and verified by hematologist who deals extensively with leukemia cases. This system is practically viable for patient database management system and text retrieval system. However image retrieval has to be improvised to retrieve images that are diagnostically relevant to meet medical experts' standard.

REFERENCES

Beksaç M, Sinan Beksaç M, Tipi BV, Duru AH, Karakas U¸ Nur çakar A (1997). An artificial intelligent diagnostic system on differential recognition of hematopoietic cells from microscopic images, cytometry (Communications in Clinical Cytometry)., 30:145-150.

Bhattacharya A (1943). On a measure of divergence between two statistical populations defined by their probability distributions, Bull. Calcutta Math. Soc., 35: 99-109.

Bishop CM (1996). Neural networks for pattern recognition, Oxford University Press Inc, United Kingdom.

Comaniciu D, Meer P, Foran D, Medl A (1998). Bimodal System for Interactive Indexing and Retrieval of Pathology Images. IEEE Proceedings on: Applications Computer Vision, USA, pp. 76-81.

Comaniciu D, Meer P (2001). The Variable Bandwidth Mean Shift and Data-Driven Scale Selection. IEEE Int. Conf. Computer Vision (ICCV'01), Vancouver, Canada., 1: 438-445.

Demir C, Yener B (2005). Automated cancer diagnosis based on histopathological images: a systematic survey, Technical report, Rensselaer Polytechnic Institute, Department of Computer Science. tr-05-09.

Giuseppe A, Vlastislav D, Michal B, Pavel Z (2005). Similarity Search: The Metric Space Approach, Springer Publishing, Computers / Data Base Management, NY, USA.

Hengen H, Spoor S, Pandit M (2002). Analysis of Blood and Bone Marrow Smears using Digital Image Processing Techniques, SPIE Medical Imaging, San Diego., 4684: 624-635.

Ian H, Frank W, Frank E (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco.

Olivier DN, Vega F (2000). Image Prototype Similarity Matching for Lymph Node Hemopathology. IEEE Computer Society, International

Conference on Pattern Recognition, Barcelona, Spain., 2: 2279-2282.

Rasband WS (2007). Image J, U.S. National Institutesof Health, Bethesda, Maryland, USA. http://rsb.info.nih.gov/ij/.

Rozenberg G (2003). Microscopic Haematology – A practical guide to laboratory, second edition, Martin Dunitz, London, United Kingdom.

Walker RF (1997). Adaptive Multi-Scale Texture Analysis with Application to Automated Cytology. PhD thesis. The University of Queensland.

Walker RF (2007). Int. J. Gynecological Cancer. January/February, p. 171: 118.

Wang C, Jing F, Zhang L, Zhang HJ (2006). Scalable search-based image annotation of personal images. International Multimedia Conference, Proceedings of the 8th ACM international workshop on Multimedia information retrieval, CA, USA.

Yan P, Yu J, Hurson AR, Potok TE (2006). Semantic-Based Information Retrieval of Biomedical Data. Symposium on Applied Computing, Proceedings of the ACM symposium on Applied computing, Dijon, France.