Content based Image Retrieval (CBIR) System for Diagnosis of Blood Related Diseases

Jignyasa Sanghavi
Assistant Professor
Department of Computer Science and Engineering
SRCOEM, Nagpur

Deepali Kayande
Assistant Professor
Department of Computer Science and Engineering
SRCOEM, Nagpur

ABSTRACT
The identification of the disease is very crucial step for curing disease. In many cases microscopic analysis of peripheral blood samples by medical practitioner is an important test in the procedures for the diagnosis of any blood related disease. Accurate diagnosis of disease is crucial for curing and controlling that disease. Earlier this process was carried out only by medical experts but now a day's automated system based on computer vision methods or image processing algorithms can speed up this operation. The expert systems are developed which are doing computerized diagnosis of various diseases using digital images of blood samples. Digital images are acquired using a digital camera connected to microscope. The presented paper shows the automatic diagnosis for three diseases i.e. Leukemia, Malaria and Sickle cell Anaemia. The system firstly segments the infected cells of leukemia and sickle cell anaemia or parasites of malaria from the blood samples and extracts the features of these cells or parasites. These features are then compared with database and accordingly classification is done and is represented in CBIR (Content Based Image Retrieval) framework.

General Terms
Image Processing, Image Retrieval System and Classification based on Feature Matching.

Keywords
Malaria, leukemia, sickle cell anaemia, expert system, CBIR (Content based image retrieval) framework, automated system, computerized diagnosis.

1. INTRODUCTION
With the increasing importance of genome project, machine vision systems have acquired more importance in medical domain. Medical fields like bioinformatics and biomedical imaging are using expert systems for their research work. Machine vision systems for the unsupervised automation usually require image processing components with exceptionally high accuracy rates. This is especially true in the biomedical domain where failures result in wrong diagnosis of disease.

1.1 Leukemia
White blood cells (WBC) or leukocytes play a significant role in the diagnosis of different diseases, and therefore, extracting information about that is valuable for hematologists. Since 1950s' projects are initiated for diagnosis of diseases using expert systems. Till now many has reached to good accuracy level in a real environment. Digital image processing techniques have helped to analyze the cells that lead to more accurate, standard, and remote disease diagnosis systems. However, there are a few complications in extracting the data from WBC due to wide variation of cells in shape, size, edge, and position. Moreover, since illumination is imbalanced, the image contrast between cell boundaries and the background varies depending on the condition during the capturing process. The microscope inspection of blood slides provides important qualitative and quantitative information concerning the presence of hematic pathologies.

From decades this operation is performed by experienced operators, which basically perform two main analyses. The first is the qualitative study of the morphology of the cells and it gives information of degenerative and tumoral pathologies such as leukemia. The second approach is quantitative and it consists of differential counting the white blood's cells.

Leukemia is a cancer of blood-forming cells in the bone marrow. These deranged, immature cells accumulate in the blood and within organs of the body. They are not able to carry out the normal functions of blood cells. The four main types of leukemia include acute lymphocytic leukemia (ALL), chronic lymphocytic leukemia (CLL), acute myelocytic leukemia (AML), and chronic myelocytic leukemia (CML).

Acute Lymphocytic Leukemia (ALL), also known as acute lymphoblastic leukemia is fatal if left untreated due to its rapid spread into the bloodstream and other vital organs, and it mainly affects young children and adults over 50. Early diagnosis of the disease is crucial for the recovery of patients, especially in the case of children. The symptoms of ALL are common also in other diseases and for this reason, the diagnosis is very difficult[1]. One of the steps in the diagnostic procedures encompasses the microscope inspection of peripheral blood. The inspection consists in the search of white cells with malformations due to the presence of a cancer. Interestingly, the morphological analysis does not require a blood sample because can be performed by using a single image. For this reason, this analysis is suitable for low-cost, homogenous accuracy, remote screening systems.

1.2 Malaria
Malaria is a disease caused by a protozoan parasites of the genus *Plasmodium*, namely Plasmodium falciparam, Plasmodium vivax, Plasmodium malariae, Plasmodium knowlesi and Plasmodium ovale transmitted to people through the bites of infected mosquitoes, is the most serious and widespread parasitic disease of humans. In the world one child dies of malaria every 30 seconds. The world malaria report 2008 describes that half of the world's population is at risk of malaria and an estimated 247 million cases led to nearly 881,000 deaths in 2006[1]. The definitive diagnosis of malaria infection is done by searching for parasites in blood slides (films) through a microscope. However, this is a routine and time consuming task. Besides a recent study on the field

shows the agreement rates among the clinical experts for the diagnosis are surprisingly low. Hence, it is very important to produce a common standard tool which is able to perform diagnosis with same ground criteria uniformly everywhere.

In peripheral blood sample visual detection and recognition of *Plasmodium species* is possible and efficient via a chemical process called (*Giemsa*) staining. The staining process slightly colorizes the red blood cells (RBCs) but highlights *Plasmodium species* parasites, white blood cells (WBC), and platelets or artifacts. The detection of *Plasmodium species* requires detection of the stained objects. However, to prevent false diagnosis the stained objects have to be analyzed further to determine if they are parasites or not. Of the five species known to infect humans, *Plasmodium falciparum* is the most virulent and contributes to the majority of deaths associated with the disease. But malaria is preventable and curable, if the patient is correctly diagnosed in early stage[1].

1.3 Sickle cell Anaemia

Sickle cell Anaemia is a genetic disease caused due to abnormality of haemoglobin. Red blood cells are round and smooth which allows them to move easily through blood vessels. The sickle cell haemoglobin causes the red blood cells to change round shape to crescent (or sickle) shape. These cells are hard and sticky, so do not easily pass through blood vessels. This disease affects approximately one in 500 African Americans and one out of 1000 to 1400 Hispanic Americans. About 72,000 people in the U.S. currently suffering from sickle cell [23].

1.4 Review of feature extraction and classification techniques of blood diseases

Elliptic Fourier descriptors were apparently overlooked in the recent revival of techniques for image content-based indexing, in spite of having the advantage of preserving topology when the Fourier series are truncated [6]. James Wang used wavelet to characterize the local texture properties within pathology images [15]. The standard procedure present in the Matlab Image Processing Toolbox: Area, Perimeter, Convex Area, Solidity, Major Axis Length, Orientation, Filled Area, Eccentricity. In addition the ratio between the cell and nucleus areas, the nucleus' "rectangularity" (ratio between the perimeter of the tightest bounding rectangle and the nucleus perimeter), the cell "circularity" (ratio between the perimeter of the tightest bounding circle and the cell perimeter), the number of lobes, and finally the solidity, area and mean gray-level intensity of the cytoplasm can be used as features [18].

Different candidate features: color histogram, Hu moments, color auto correlogram, and a relative shape measurements vector are used to investigate their individual performances and then search for a higher combined feature performance. The histogram is a widely used descriptor which is simple to compute and gives adequate information about the color distribution. Hu's moment invariants are derived from algebraic combinations of the first 3 orders of normalized central moments. They are also rotation and scale invariant while providing spatial information. The third feature, what we call the relative shape measurements vector, is formed of simple measurements to represent the object shape [11]. Similarity of images in content based image retrieval system is evaluated based on four image feature types: color histogram, image texture, Fourier coefficients and wavelet coefficients using the vector dot product as a distance metric [22].

Various classification approaches adopted by researchers in the past to classify blood diseases are discussed in this section. Y. Hirimutugoda implemented two back propagation Artificial Neural Network models (3 layers and 4 layers) together with image analysis techniques to evaluate the accuracy of the classification in the recognition of medical image patterns associated with morphological features of erythrocytes in the blood. The three layers Artificial Neural Network (ANN) architecture had the best performance with an error of 2.74545e-005 and 86.54% correct recognition rate. The trained three layer ANN acts as a final detection classifier to determine diseases [9]. To classify the stained objects a distance weighted K-nearest neighbour classifier is implemented [11]. J. Soni implemented a two stage tree classifier that distinguishes between true and false positives and then diagnoses the species of malaria infection [8]. In experiments of [3] the classification was carried out using instance based classifiers, decision trees, regression functions as well as meta classifiers available in (Weka 2009). Some of these are: K-Nearest Neighbour, Random Forest, Simple Logistic, SMO and Random Committee which were chosen because they were able to obtain best results for leukemia best subtypes classification in the work of (Galindo 2008) [3]. Ten-fold cross validated classification was implemented to provide more realistic estimation of the system performance [19]. Feed-forward neural networks with log-sigmoidal activation function (FFNN) and with two hidden layers have been created by ranging the number of the hidden units from 2 to 50**.** Then Levemberg Marquard training method was used with Bayesian regularization present in the Matlab Neural Network Toolbox [18].

1.5 CBIR Model

Content Based Image Retrieval (CBIR) has emerged in the early 1990s. The principal aim is to represent each image as a feature vector and to measure the similarity between the image and image database and to retrieve similar digital images based on features and not on textual annotations. In CBIR query is given in the form of image. Then image database is searched through all images in order to find those with the most similar indices which are returned as the images most alike to the query image. Figure 1 presents the general Architecture of CBIR systems proposed in [29].

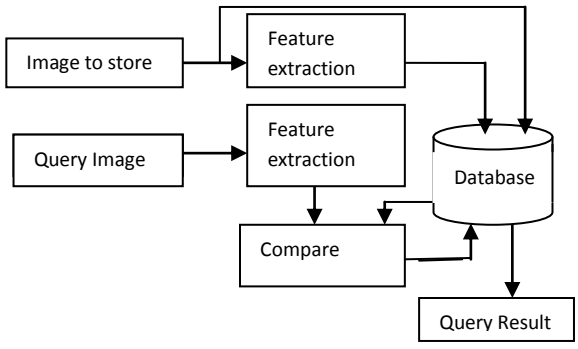

Fig 1: General Architecture of CBIR systems

1.6 Medical CBIR

Despite various CBIR developments, medical images in different fields are very particular and require a specific design of CBIR systems. There exists a large number of medical image acquisition devices among which computed tomography scanners (CT), magnetic resonance imagers (MRI), ultrasound probes (US) and nuclear imagers are the

most widely used. They provide images with very different properties in terms of resolution, contrast, and signal to noise ratio. They are highly specialized and they produce images giving very different information on the human body anatomy and physiology. A.Kak and C. Pavlopoulou notice that medical image retrieval must often be processed according to pathology bearing regions which are precisely delimited on the images and could not automatically detected in the general case[30].Moreover, low level features like colour, texture or shape are not sufficient to describe medical images [31]. As a consequence, medical CBIRs require a high level of content understanding and interpretation of images, which implies their automatic segmentation [32]. Finally, a high level of query completion and accuracy is required by such systems to make them reliable from a clinical point of view [33].

The system we proposed first segments the infected cell or parasite from the other blood cells then it extracts the morphological features from that cell, compare the features with the database and finally classify them using CBIR framework.

The rest of the paper is organized as follows: Section 2 discusses the proposed system for classification of blood diseases in detail following with the results of each step, section 3 presents the experimental results on blood image classification and section 4 concludes the paper.

2. THE PROPOSED SYSTEM

Fig. 2 shows the proposed CBIR framework for diagnosis of blood diseases. The proposed system consists of Image database of peripheral blood images, from which the relevant features are extracted and stored in feature database.

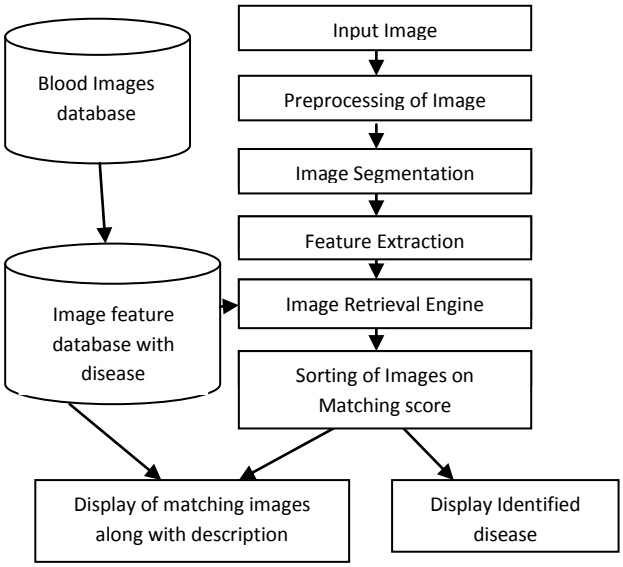

Fig. 2: The Proposed CBIR framework for disease classification

The expert knowledge from pathology experts in the form of disease class for each peripheral blood image is also recorded in the feature database itself. This classification task for each image is done offline by the pathology experts.

2.1 Image Pre-processing

In this module first the input image is resized to 256 x 256 image. This resized image is then converted to the gray scale image. Then by applying appropriate limits contrast stretching is applied to the gray scale image so that the infected cells or parasites get darker and unwanted background becomes lighter.

2.2 Image Segmentation

In this step the unwanted background of the image is removed by applying proper thresholding to contrast stretched image and converting it to binary image. In binary image some small spots or points are observed so to remove them Opening is performed followed with closing on the binary image. For evaluating the Fourier descriptor we are complementing the resultant binary image.

2.3 Feature Extraction

In this step the features are extracted from the region of interest (ROI).In this project the four features are extracted:

- Area of object.
- Perimeter of object.
- Fourier descriptor.
- Local Binary Pattern (LBP).

From this we are calculating metric of the shape by following formula:

$$m = \frac{Area}{Perimeter}$$

Fourier Descriptor is used as shape descriptor and its plot is used for matching. Local Binary Pattern is used for texture analysis, LBP variance points are used in matching process.

2.4 Matching Techniques

In this step we are finding images from database matching with query image using following matching techniques:

- Euclidean Distance
- Fourier Descriptor matching
- Subtraction.

The metric calculated is simple variable so direct subtraction with each value of database is performed. The LBP Variance points is an array of 59 elements so Euclidean distance is calculated with values of database. The Fourier Descriptor matching is used to calculate the matching percentage of Fourier Descriptor Plot of query image with that of database. These features are given proper weight age so as to get accurate results. The weight age to the features is shown in table1.After calculating the total percentage; the threshold matching score is set to 60%. Images having matching score greater than 60% are only retrieved from this system.

Table 1: Parameter Weight age

Texture	Shape descriptor	Area/Perimeter
0.7	0.1	0.2

2.5 Retrieval Technique

Content based framework is used for retrieving the matching images as well as for classification of the disease. Along with the images retrieved their matching percentage score is also displayed. Images are displayed in descending order of their matching percentage score.

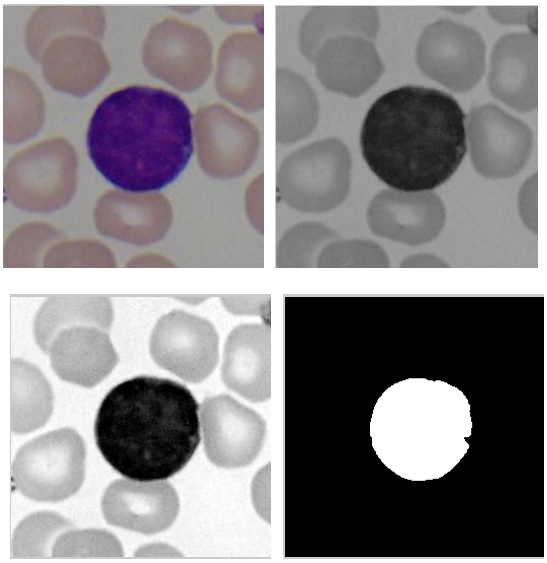

Fig 3: (a) Original image (b) Grayscale image (c) Contrast Stretched image (d) Binary image

Fig 4: Fourier Descriptor of above image

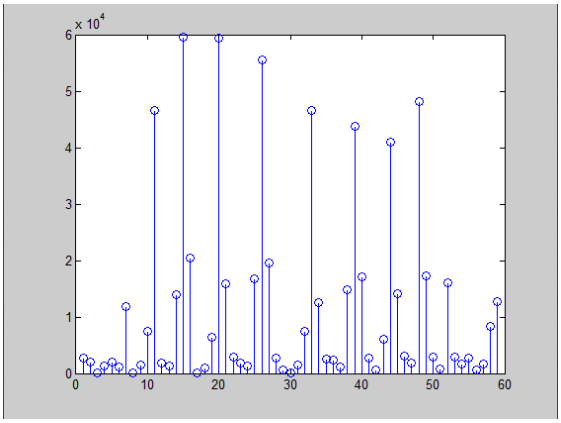

Fig 5 : LBP variance graph

The features extracted from query image are matched with feature database. The matching percentage score of each image of database with query image is calculated. The threshold of 60% is set for retrieval. The images matching with query image with high matching percentage score are displayed in output window in descending order of their matching score with their matching score as shown in figure 6.

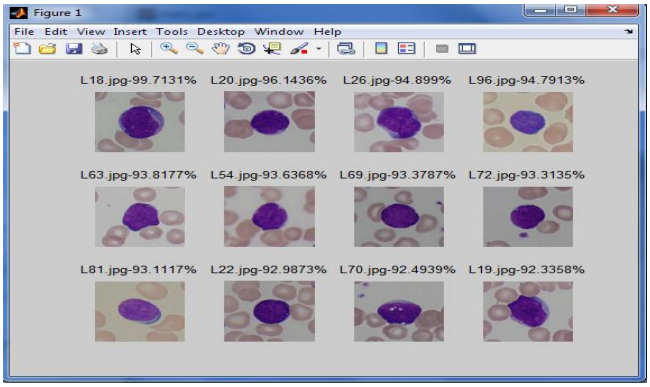

Fig 6 : Output of CBIR system of pathology images

3. EXPERIMENTAL RESULTS

This system is implemented for the sample images of leukemia, malaria and sickle cell anaemia. The dataset consist of 260 images of leukemia, 144 images of malaria and 23 images of sickle cell anaemia. Using this system we have secured approximately 95% accuracy for Leukemia images, approximately 88% accuracy for malaria and 75% accuracy for sickle cell anaemia images. These variations are due to unavailability of digital images for the disease.

4. CONCLUSION

This system presents a use of CBIR (Content Based Image Retrieval) framework for classification of few blood related diseases using digital microscopic blood film images. This system is developed for three diseases i.e. Leukemia, Malaria and Sickle Cell Anaemia. It identifies and proposes new research challenges in the medical domain. It will speed up the process of disease diagnosis. This system is not only useful to medical practitioners as well as to the medical students in their studies. This system is useful in rural areas. In the future the coverage of the problem can be extended to more diseases provided digital microscopic blood film images are available for more diseases. This system can also be connected to patient's database. So when we are retrieving images we can have patients data also retrieved with the image of blood sample, who suffered from that disease. The history of previous patients can be used for analysis and to take correct decision in treatment.

5. ACKNOWLEDGEMENT

We wish to thank Dr Arvind Malaviya from Malaviya Pathology Nagpur, India for their profitable co-operation and encouragement and Dr. Fabio Scotti of Department of Information Technology from Crema, Italy for providing us dataset of Leukemia.

6. REFERENCES

[1] World Health Organization. What is Malaria? and What is Leukemia? http://apps.who.int/gb/ebwha/pdf_files/EB118/B118_5-en.pdf.

[2] R.Labati,et al, "ALL-IDB:The Acute Lymphoblastic Leukemia Image Database for image processing",IEEE conference on Image Processing,2011.

[3] C.Reta,et al, "Segmentation of Bone Marrow Cell Images for Morphological Classification of Acute

Leukemia",Proceedings of 23rd International Florida Artificial Intelligence conference,2010.

[4] F.Scotti, "Robust Segmentation and Measurements Techniques of White Cells in Blood Microscope Images", IMTC conference Italy, 2006.

[5] F.Scotti, "Automatic Morphological Analysis for Acute Leukemia Identification in Peripheral Blood Microscope Images", IEEE conference on Computational Intelligence,2005.

[6] D.Foran, et al, "Computer –Assisted Discrimination Among Malignant Lymphomas and Leukemia Using Immunophenotyping,Intelligent Image Repositories and Telemicroscopy",IEEE Transactions on Information Technology in Biomedicine,vol 4,2000.

[7] D.Anggraini, et al, "Automated Status Identification of Microscopic Images obtained from Malaria Thin Blood Smears" IEEE conference Indonesia,2011.

[8] J.Soni, "Advanced Image Analysis based system for automatic Detection of Malarial Parasite in Blood Images using SUSAN approach" IJEST Vol.3 No.6, 2011.

[9] Y.Hirimutugoda, et al, "Image Analysis System for Detection of Red Cell Disorders using Artificial Neural Network" Sri Lanka Journal of Bio-Medical Informatics,2010.

[10] F.Boray Tek, et al, "Computer vision for microscopy diagnosis of malaria" Malaria Journal 2009.

[11] F.Boray Tek, et al, "Malaria Parasite Detection in Peripheral Blood Images" Proceedings of Medical Image understanding and Analysis conference,2006.

[12] J.Gim, et al, "A novel framework for white blood cell segmentation based on stepwise rules and morphological features",Proceedings of SPIE,2011.

[13] B.Ko, et al, "Microscopic cell Nuclei Segmentation Based on Adaptive Attention Window" Proceedings of Journal of Digital Imaging vol 22,June,2009.

[14] F.sadeghian, et al, "A Framework for White Blood Cell Segmentation in Microscopic Blood Images Using Digital Image Processing",Proceedings of Biological Procedures Online,vol 11,2009.

[15] J.Wang , et al, "Pathfinder: Multiresolution Region-Based Searching of Pathology Images using IRM",Proceedings of the AMIA Symposium,2000

[16] C.Carson, et al, "Blobworld: A System for Region Based Image Indexing and Retrieval ",Springer,Visual Information Systems,1999.

[17] Rafeal C.Gonzalez,Richard E.Woods,Digital Image Processing,2nd Edition,Prentice Hall,2006.

[18] V.Piuri and Fabio Scotti, "Morphological Classification of Blood Leucocytes by Microscope Images" IEEE conference USA,2004.

[19] D.Comaniciu et al, "Bimodal System for Interactive Indexing and Retrieval of Pathology Images",0-8186-8606-5/98 IEEE,1998.

[20] D.Comaniciu et al, "Shape-Based Image Indexing and Retrieval for Diagnostic Pathology", IEEE conference on Pattern Recognition,1998.

[21] L.Dorini et al, "White blood cell segmentation using morphological operators and scale-space analysis", IEEE conference on Computer Graphics and Image Segmentation,2007.

[22] Lei Zheng et al, "Design and Analysis of a Content-Based Pathology Image Retrieval System"IEEE Transactions on Information Technology in Biomedicine, vol.7, no.4, December 2003.

[23] www.netofcare.org/content/pdf/6-spec_illness-sicklecell.pdf

[24] http://dti.unimi.it/fscotti/all

[25] http://dpd.cdc.gov/DPDx/HTML/ImageLibrary/Malaria_il.htm

[26] http://www.emedicinehealth.com/leukemia/article_em.htm

[27] J.Sanghavi et al,"Morphological Classification of few blood related disease using peripheral Blood microscopic images"Proceeding of ITAC,2012.

[28] Topi Maenpaa & Matti Pietikainen,"Texture Analysis with Local Binary Patterns", WSPC Review volume, May 13, 2004.

[29] T. Lehmann, B. Wein, J. Dahmen, J. Bredno, F. Vogelsang and M. Kohnen, "Content-Based Image Retrieval in Medical Applications : A Novel Multi-Step Approach", *In International Society for Optical Engineering (SPIE), volume 3972(32), pages.*

[30] A.Kak and C. Pavlopoulou. "Content-Based Image Retrieval from Large Medical Databases. In 3D Data Processing," *Visualization, Transmission, Padova, Italy, June 2002*

[31] F. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, "Fast and effective retrieval of medical tumor shapes". *IEEE Trans. on Knowledge and Data Engineering,10(6):889–904, 1998.*

[32] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler. "A review of content-based image retrieval systems in medical applications – clinical benefits and future directions." *International Journal of Medical Informatics, 73(1):1{23, Feb. 2004*.

[33] H. D. Tagare, C. C. Ja_e, and J. Duncan. Medical Image Databases: " A Content-based Retrieval Approach". *J Am Med Inform Assoc, 4(3):184,198,1997*.