2009 Third Asia International Conference on Modelling & Simulation

Content-based Image Retrieval for Blood Cells

Mohammad Reza Zare, Raja Noor Ainon, Woo Chaw Seng
Faculty of Computer Science and Information Technology
University of Malaya
Kuala Lumpur, Malaysia
mrezazare2004@yahoo.com , ainon@um.edu.my, cswoo@um.edu.my

Abstract— **The rapid development of technologies and steadily growing amounts of digital information highlight the need of developing an accessing system. Content-based image indexing and retrieval has been an important research area in computer science for the last few decades. The approaches of content-based image retrieval using low level features such as colour, shape and texture are investigated to create a prototype that perceives blood cell images similar to a human. The histogram of red, green, and blue colour components is analyzed. The wavelet decomposition is also used to analyze texture. In addition, morphological operations such as opening and closing are applied to analyze object shape. Lastly, colour, texture, and shape in image retrieval are integrated in order to increase the retrieval accuracy. Experimental results using four different classes of 150 blood cell images showed 95.68% of retrieval accuracy.**

Keywords: Text-based Image Retrieval (TBIR), Content-based Image Retrieval (CBIR), Blood cell images

I. INTRODUCTION

In earliest times the knowledge of a community was concentrated in the minds of the elders and sages. Consequently searching for information meant asking these people. But the world we live in is becoming inherently interconnected and digital. Today the knowledge and information mankind has collected exceed the mental capacity of any single human mind and forms a vast amount of data. Moreover the rapid development of technologies and computing hardware makes the digital acquisition of information to be more in demand and popular in recent years.

Therefore many digital images are being captured and stored such as medical images, architectural and engineering images, advertising, design and fashion images, etc., and as a result large image databases are being created and used in many applications. However, the focus of our study is on medical images in this work. A large number of medical images in digital format are generated by hospitals and medical institutions everyday. Consequently, how to make use of this huge amount of images effectively becomes a challenging problem [1].

The most common approach that had been used previously for image retrieval was Text Based Image Retrieval (TBIR). In TBIR, all medical images used to be annotated with text which would enable them to be accessed by a text-based searching engine. This text file is used to store the descriptive captions or keywords for each and every image. One of the problems of this method is that the keyword is man made and is subjective to an individual, so the differences in human perception when describing the image may cause inaccuracies when it comes to retrieval. It is also difficult to have an identical label or keyword for two different images which contain almost similar objects as well as being very expensive and time consuming to implement manual annotation [2]. The other problem is that the contents of some medical images are difficult to be exactly described in textual form due to the irregular organic shape.

These problems limit the feasibility of text-based search for medical image retrieval [3]. Image databases are being used due to the large amount of images that are generated by various devices and applications; and this brings about a need for an efficient and automatic procedure for retrieving images from these databases, especially for the blood cell image database where the content of images is very similar and therefore it becomes imperative to use very precise descriptors.

In order to overcome this problem, an alternative approach was proposed for retrieving the images from a database, called Content Based Image Retrieval (CBIR). Content based image retrieval is a retrieval of images based on visual features such as texture, shape and colour. In CBIR, the features from images are extracted and then the images are compared with one another based on those extracted features. Those images which have similar features would have similar content as well.

Novel methods based on image retrieval to classify and retrieve medical images from highly similar image databases have been proposed in [4, 5, and 6]. In this paper we propose an approach using low-level features such as colour, shape and texture to create a prototype that perceives blood cell images similar to a human.

II. PROPOSED METHOD

The primary goal of a typical CBIR system is to find the nearest image to the query image. However, due to the existence of a large number of medical image acquisition devices, medical images are distinct and require a specific design of CBIR systems. The goals of medical information systems have been defined to deliver the needed information at the right time, the right place to the right person in order to improve the quality and efficiency of care processes.

978-0-7695-3648-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AMS.2009.103

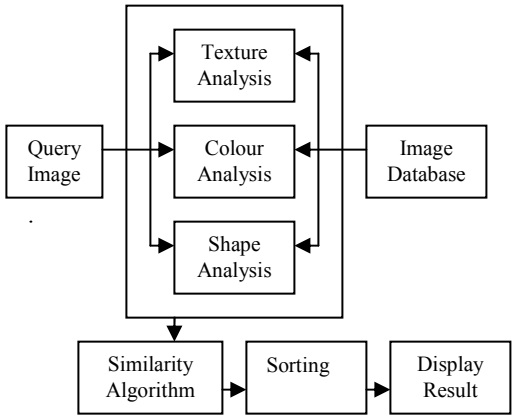

Figure 1. Framework of the proposed System

In the medical domain, images from the same disease class as the query image must be retrieved in order to help the doctor in diagnosis. The images in the medical database are labelled by a specialist to ensure that they are less subjective than those of the generic CBIR [5]. Figure 1 represents the framework of the proposed CBIR system.

This level of retrieval is based on the primitive features. The following are some of the primitive features such as colour, texture, shape or the spatial location of image element.

A. Colour Analysis

Colour is one of the most important features that make the image recognition possible by human. It is a property that depends on the reflection of light to the eye and the processing of that information in the brain.

Colour will be used everyday to differentiate objects, places, etc. where colours are defined in three dimensional colour spaces such as ***RGB*** (Red, Green, and Blue), ***HSV*** (Hue, Saturation, and Value) or ***HSB*** (Hue, Saturation, and Brightness).

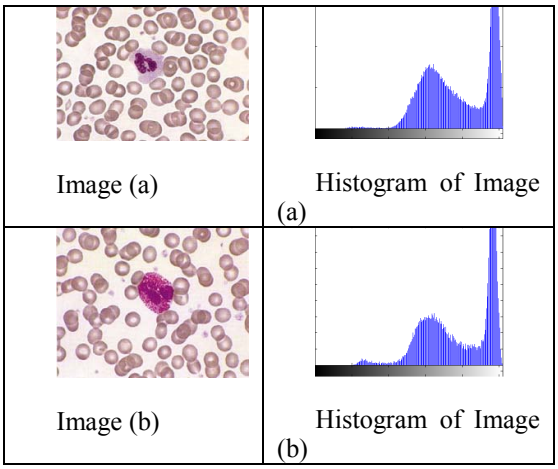

Figure 2. Sample Blood Cell Images and Its histograms

Bhattacharya coefficient between the images in Fig.2(a) and Fig.2(b) is 0.984.

Most image formats use the RGB colour space to store information [7]. The RGB colour space is defined as a unit cube with red, green, and blue axes. Thus, a vector with three co-ordinates represents the colour in this space which represents black when all of them set to zeros and represents white when all three coordinates are set to 1. In CBIR, colour histogram is the main method to represent the colour information of the image. A colour histogram is a type of bar graph, where each bar represents a particular colour of the colour space being used.

We have calculated the colour histograms of query image and images in a database and put them into two different vector and compare them using Bhattacharya coefficient. The **Bhattacharya coefficient** is an approximate measurement of the amount of overlap between two statistical samples. The coefficient can be used to determine the relative closeness of the two samples being considered.

Calculating the Bhattacharya coefficient involves a rudimentary form of integration of the overlap of the two samples. The interval of the values of the two samples is split into a chosen number of partitions, and the number of members of each sample in each partition is used in the following formula:

$$\text{Bhattacharya} = \sum_{i=1}^{n} \sqrt{(\sum ai \bullet \sum bi)} \qquad (1)$$

Where considering the samples ***a*** and ***b***, ***n*** is the number of partitions, and a_i, b_i are the number of members of samples ***a*** and ***b*** in the i^{th} partition. The Bhattacharya coefficient will range from 0 to 1 where 1 represents the completely similar image and 0 indicates that there is no similarity in two images. The concept of normalization will be used in Bhattacharya coefficient; **normalization** is a process that changes the range of pixel intensity values. The purpose is to bring the image with a different intensity values into a range that is more familiar and similar to the senses which in this case, the ranges is brought to 0 to 1.

B. Texture Analysis

A method called the pyramid-structured wavelet transform for texture classification is used. It decomposes sub-signals in the low frequency channels recursively. It is mainly trivial for textures with dominant frequency channels. For this reason, it is mostly suitable for signals consisting of components with information concentrated in lower frequency channels. Since most of the information exists in lower sub band of the image due to the natural image properties, the pyramid-structured wavelet transform is highly sufficient.

Using the pyramid-structured wavelet transform, the texture image is decomposed into four sub images, in low-low, low-high, high-low and high-high sub-bands. At this point, the energy level of each sub-band is calculated which is the first level decomposition. In this study, fifth level decomposition is obtained by using the low-low sub-band for further decomposition. The reason for this is the basic

assumption that the energy of an image is concentrated in the low-low band. For this reason the wavelet function used is the Daubechies wavelet. [8]

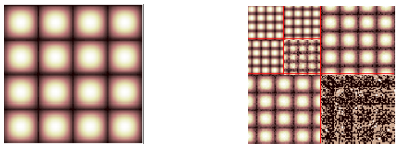

Figure 3. Original and Decomposed Image

C. *Cell Geometric Analysis*

Based on the domain in this project which is blood cell images, the number of round objects in the image needs to be determined; to achieve this, an image will be converted to black and white in order to prepare for boundary tracing bwboundaries function in MATLAB.

Then morphological operators such as opening and closing will be used to remove the small connected objects which do not belong to the objects of interest. The result of area and perimeter of an object inside each image will be used to form a simple metric indicating the roundness of an object using the following formula:

$$\text{Metric} = \frac{4\pi \times \text{area}}{\text{perimeter}^2} \qquad (2)$$

This metric is equal to one only for a circle and it is less than one for any other shape. The discrimination process can be controlled by setting an appropriate threshold. In this work, the threshold of 0.50 has been used since all the objects or bubbles in blood cell images are not completely round. The formula (2) is used to calculate the roundness of each object in Figure 4.

0.90 0.95 0.91

Figure 4. Examples of Roundness Metric

D. *Similarity Algorithm and Sorting*

In colour analysis, the value of colour histogram of query image will be compared with the value of colour histogram of an image in database and the result is a number ranging from zero to one based on the normalization algorithm. This value and its respective image will be saved in a vector if it is greater than the value of threshold which is 0.97. This process will be repeated for all images in database. At the end, the vector will be sorted in ascending order and the top ten values will be selected and their respective image will be displayed as a retrieved images.

As for texture analysis, the Euclidean distance is calculated between the query image and every image in the database and the value will be compared with a threshold value which is 1.7. If it is smaller than the threshold value, the value and its respective image will be saved in the vector. This process will be repeated for all images in database.

As for cell geometric analysis, once the percentage of oval objects in the query image is calculated, its value will be compared with all the value of percentage of oval objects in all the images in database. Then those values which are close to query image's value by ***±3*** will be saved in the vector as well as its respective image and finally the top ten closest value and images will be retrieved.

III. EXPERIMENTAL RESULTS

In our classification system, the ground truth database is made of 150 blood cell images with four different classifications. In Table 1, the precision retrieval result of 16 different query images from different categories as well as the average precision retrieval result is presented. The accuracy precision of each features is calculated for each image separately, then the value of related features of all the 16 query images are added and the average precision retrieval is computed as it can be seen in the table.

In general, it is difficult to compare any two retrieval systems in the image retrieval domain. For medical image retrieval system, the evaluation issue is almost non existent in most of the published papers. Due to the strong noise in most of the medical images [9] as well as the existing similarities in the content of the images, it becomes imperative to use very precise descriptor.

However, In order to increase the accuracy of retrieval result in the proposed system, the result of colour, texture and cell geometric are combined so that only images which are common in all the above three feature extraction will be shown as final result. The advantages of this system are high accuracy and precision of 95.68**%** as well as simplicity of the algorithm.

TABLE I. PRECISION RESULT FOR 16 QUERY IMAGES

Retrieval Methods	Colour Extraction	Texture Extraction	Cell Geometric Extraction	Final Retrieval Result
Query1	64.5	70	90	100
Query2	55	43	35	60
Query3	60	87.5	80	85
Query4	64	57	86	86
Query5	55	100	90	100
Query6	80	100	80	100
Query7	67	100	80	100
Query8	100	100	100	100
Query9	100	68	82	100
Query10	70	40	80	100
Query11	65	100	80	100
Query12	70	72	90	100
Query13	75	100	80	100
Query14	80	50	70	100
Query15	100	80	90	100
Query16	70	80	100	100
Average	**73.46**	**77.96**	**82.06**	**95.68**

Cell geometric extraction performs the best with average precision of **82.06%** while texture and colour extraction has the precision of **77.96%** and **73.46%** respectively. As expected, cell geometric based methods are significantly slower than colour and texture based methods. This could pose problem when it goes to larger database.

IV. CONCLUSION

The vivid growth in the sizes of image databases highlights the need of developing an effective and efficient retrieval system. This development started with retrieving images using textual annotation called Text Based Image Retrieval (TBIR) but later introduced image retrieval based on content which is known as Content Based Image Retrieval (CBIR)**.** Content based image indexing and retrieval has been one of the most important research areas in computer science for the last decade. The large number of research publication has been done for CBIR especially in the field of medical domain.

This work studies the weaknesses of text based image retrieval in medical domain. In TBIR, images would be retrieved by text based search engine which is suffering from few major drawbacks such as time consuming and expensive, it is subjective to an individual and it fails to deal with inconsistency of subjective perception. Another achievement is to develop a prototype for retrieving medical images similar to a human.

This work also investigates the approaches of current content based image retrieval (CBIR) based on the low level features such as colour, shape and texture analysis. In this work, the modified algorithm to combine the three indexing methods to assist and accelerate the feature extractions such as colour, texture and cell geometric were proposed. The retrieval accuracy of 95.68% is good by adjusting the threshold value and can be enhanced to achieve excellency by further improvement on the algorithm.

V. FUTURE WORK

The CBIR technology is exciting but immature and even though it overcomes the limitation and shortage of text based system to a certain degree, some limitations and drawbacks still existed which is the gap between high level feature and low level features. In addition, due to the lack of human perception, only low level features are not suitable enough in image retrieval. This shortcoming of CBIR system will be analyzed in future work by integrating of low level features with high level features. The capability of the system can be improved by using the multi level search to narrow down the retrieval result.

The histogram captures only the colour distribution and it does not include any spatial correlation between individual pixels which it may cause to have limited discriminative power. Therefore, colour correlogram can be explored to minimize this shortcoming. At the same time, wavelet-based correlogram method can be investigated.

REFERENCES

[1] Lehmann, T.M., Wien,B., Dahmen,J., Bredno,J., Vogelsang,F. & Kohnen, M. (2000) Content based image retrieval in medical applications: a novel multi step approach. *International Society for Optical Engineering (SPIE),* Vol. 3972, pp. 312-320

[2] Zare, M. R., Woo,C.S. & Norfizlina,J. (2008) Comparative Analysis of Image Retrieval approaches. *International Conference of Biomedical Engineering*, Kuala Lumpur, Malaysia, Vol 21, pp 847-850

[3] Eakins, J.P. (2002) Towards intelligent image retrieval. *Pattern Recognition*, Vol. 35, Issue. 1, pp. 3-14

[4] Angulo,J. & Serra,J. (2002) Morphological colour size distributions for Image classification and retrieval. *Proceeding of ACIVS (Advanced Concept for Intelligent Vision Systems),* Ghent, Belgium

[5] Cecilia, D.R., Andrew,D. & Shahid,K. (2002) Analysis of infected blood cell images using morphological operators. *Image and Vision Computing 20*, pp.133-146

[6] Pan, C., Yan, X. & Zheng, C. (2006) Recognition of blood and bone marrow cells using kernel-based image retrieval. *International Journal of computer Science and Network security,* 6**,** 7.

[7] Hengen,H., Spoor,S. & Pandit, M. (2002) 'Analysis of Blood and Bone Marrow Smears using Digital Image Processing Techniques*'. SPIE Medical Imaging*, San Diego, Vol. 4684, pp. 624-635.

[8] Siddique, S. (2002) A Wavelet Based Technique for Analysis and Classification of Texture Images. Ottawa, Carleton University.

[9] Glatard, T., Montagnat, J. & E.Magnin, I. (2004) Texture based medical image indexing and retrieval: Application to cardiac imaging. *In proceeding of the ACM SIGMM International Workshop on Multimedia Information Retrieval.*