A Novel Approach for Content Based Microscopic Image Retrieval System Using Decision Tree Algorithm

B.Ramasubramanian, G.Prabhakar, S. Murugeswari

Abstract— A medical image may represent the symptoms of a specific disease. The collection of these images can be a useful source for physicians and researchers. Many different types of medical images are produced and collected in various medical centres every day. Hence a system is needed that can efficiently retrieve medical images representing a particular disease. Content Based Image Retrieval is a method for retrieving images based on similarity of their visual content. Unlike Traditional text based methods, this approach does not need time consuming and erroneous annotation processes. In our approach, Multitier Content Based Image Retrieval (CBIR) system for microscopic images having more than one disease is designed. Firstly, the input image is given as a query to the system. The features based on colour and texture is extracted. In the first tier, the images are classified by recursive SVM classifier with the help of extracted features. In the next tier, the similar images are retrieved using Decision tree algorithm. The retrieval performance of this method can be tested using medical image database and measured by finding precision and recall.

Index Terms— CBIR, GLCM, Recursive SVM, Decision Tree Algorithm.

1 Introduction

NOW a day's medical imaging is employed in the diagnosis of many diseases. Every day, large volumes of different types of medical images such as MRI, ultrasound, radiology, etc, are produced in different medical centres. These images are informative and valuable. A system that could organize and retrieve the medical images is very useful in diagnosis, training, and research purposes.

Text-based image retrieval, the first method available, is the typical and traditional method for retrieving images. In this method, images are annotated by keywords and retrieving is performed through keywords as indices of images. This method however, has many significant disadvantages including manual image annotation which is a labour intensive and time consuming process. Again, it causes errors because each word may have several meanings depending on the context. Therefore various methods and algorithms have been presented for automatic image annotation. however, since those methods describe images just with keywords, they also have some problems noted earlier.

The second method is content-based image retrieval (CBIR) which uses visual content of images to search amongst large databases. The main idea here is the extraction of visual features from an image, and stores them as an index of that. Retrieving is carried out later based on those visual features and a similarity measurement scheme i.e., in this method the user gives an image to the system as an input, and the system compares its visual contents with images in the database, and then retrieves the most similar images.

Although several CBIR projects exist for radiology [8]–[10] and several other projects are underway, there is an acute need for a comprehensive and flexible CBIR system for microscopic images with direct implications for the field of pathology and cancer research. Microscopic images present novel challenges because they 1) are large in size 2) demonstrate high degree of visual variation due to large variation in preparation (e.g., staining, thickness), and 3) show huge biological variation. Therefore, a well-designed CBIR system for microscopic images can be extremely useful resource for cancer research, diagnosis, prognosis, treatment, and teaching. In other words, such a system can 1) assist pathologists in their diagnosis and prognosis, 2) potentially help to reduce inter- and intrareader variability in clinical practice for the diseases, especially those with complicated classification, 3) help cancer researchers in better understanding of cancer development, treatment monitoring, and clinical trials, and 4) train future generation of researchers by providing consistent, relevant and always available support and assistance. In this paper, we describe the design and the development of a multitiered CBIR system for microscopic images from a reference database that contains more than one disease.

- *B.Ramasubramanian is currently working as an Assistant Professor in Syed Ammal Engineering College, Ramanathapuram, India. E-mail: ramatech87@gmail.com*
- *G.Prabhakar is currently working as an Assistant Professor in Syed Ammal Engineering College, Ramanathapuram, India. E-mail: gprabhakar2488@gmail.com*
- *S.Murugeswari is currently working as an Assosciate Professor in Syed Ammal Engineering College, Ramanathapuram, India.*

International Journal of Scientific & Engineering Research Volume 4, Issue ŀ)ÜÕÌ-2013
ISSN 2229-5518

To provide a motivating example and to test the ideas developed in this paper, images in our reference database include sample regions cropped from digitized hematoxylin and eosin (H&E) stained whole slides. Neuroblastoma (NB) and follicular lymphoma (FL) tissue images have been collected as part of our ongoing projects for both diseases. The input images to our system are digitized using a Scope XT digitizer (Aperio, San Diego, CA) at 40× magnification. FL tissue slides were collected from the Department of Pathology, The Ohio State University, in accordance with an Institutional Review Board (IRB) approved protocol. NB whole-slide tissue samples were collected from the Children's Oncology Group slides with an IRB-approved protocol. According to the recent medical statistics, FL accounts for 20%–25% of non-Hodgkin lymphomas in the US [11], [12] and affects predominantly adults, particularly the middle aged and elderly. FL cases are stratified to three histological grades from low- to high-risk category as follows: Grade I, Grade II, and Grade III. NB is the most common extracranial solid cancer in childhood and in infancy. According to the International Neuroblastoma Classification System, NB tissues are mainly divided into two subtypes such as stroma rich (SR) or stroma poor (SP) based on the degree of Schwannian stroma development [13]. Additionally, SP tissue has three subtypes such as undifferentiated (UD), poorly differentiated (PD), and differentiating (D). These subcategories as well as the mitosis karyorrhexis index are used for prognostication.

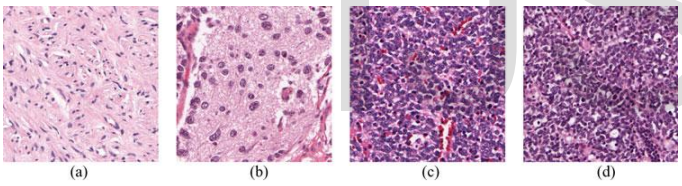

Fig. 1. Sample NB images associated with the four grades. (a) SR. (b) D. (c) PD. (d) UD

The rest of the paper is organized as follows. Section II presents related works on CBIR methods for medical images. Section III explains the preprocessing and features extracted from the database images. Classification and retrieval are discussed in section IV. .Database description and results of the experiments are discussed in Section V. Conclusions are drawn in Section VI.

2 Related Work

Quellec et al. [3] have proposed a content-based image retrieval method for diagnosis aids in medical fields. They are applied the method to two medical datasets (retinal and mammography images) and one general dataset (face images). They employed Wavelet transform in order to extract general features from the images. Then with some tunable parameters, their method will be capable to be applied for any pathology and modality.

Andaloussi et al. [4] have proposed a content-based image retrieval method for diagnosing aids. Their method extracts general features from images. They used Bidimensional Empirical Mode Decomposition (BEMD) to decompose images and used generalized Gaussian function to extract features. Tobin et al. [5] have presented a method for diagnosing retinal pathologies using a content-based image retrieval method. They employed the retrieval of similar images to estimate a posterior probability. They also used these probabilities for further diagnoses.

In another study [20], expectation-maximization algorithm was used to generate clusters of block-based low-level features extracted from radiographic images. Then, the similarity between two clusters was estimated as a function of the similarity of both their structures and the measure components. Pourghassem and Ghassemian [21] proposed a two-level hierarchical medical image classification method. The first level was used to classify the images into the merged and nonmerged classes. They tested their algorithm on medical X-ray images of 40 classes. Although this is a two-level hierarchical classification, it is different from our approach because only the merged classes were evaluated in the second level to be classified with multilayer perceptron (MLP) classifiers into 1 of 40 classes.

A few CBIR systems for the microscopic images have been developed for clinical use [6], [7], [17], [18], [19]. Mehta et al. designed a region-specific retrieval system based on sub image query search on whole-slide images by extracting scale invariant features on the detected points of interests and 80% of match was achieved with the manual search for prostate H&E images [19] in the top five searches. In another study, image-level retrieval of four special types of skin cancer [18] was performed by constructing a visual word dictionary using a bag-of-features approach in order to represent a relationship between visual patterns and semantic concepts. A prefiltering approach [9] was proposed to reduce the search space of query images by categorizing the images using multiclass support vector machines (SVMs) and fuzzy c-mean clustering.

3 Preprocessing & Feature Extraction

3.1 Pre-processing

Colour fundus images often show important lighting variation, poor contrast and noise. In order to reduce this imperfection [11] and generate images more suitable for extracting the pixel features in the classification process, a pre-processing comprising the following step is applied. 1) RGB to HSI conversion 2) Median Filtering 3) Contrast Limited Adaptive Histogram Equalization (CLAHE).

3.1.1 RGB to HSI Conversion

The input retinal images in RGB Colour space are converted to HSI colour space. The noise in the images are due to the uneven distribution of the intensity(I) component.

International Journal of Scientific & Engineering Research Volume 4, Issue ł()ÜÕÌ-2013
ISSN 2229-5518

3.1.2 Median Filtering

In order to uniformly distribute the intensity throughout the image, the I-component of HIS colour space is extracted and filtered out through a 3X3 median filter.

3.1.3 Contrast Limited Adaptive Histogram Equalization (CLAHE)

The contrast limited adaptive histogram equalization is applied on the filtered I-component of the image [12]. The histogram equalized I component is combined with HS component and transformed back to the original RGB colour space.

3.2 Feature Extraction

In this section, we will explain the feature extraction techniques based on color and texture

3.2.1 Color Feature extraction based on CCM

Assuming color image is divided into N × N image sub-block, for anyone image sub-block

T(i, j) (1 = i = N,1 = j = N) ,

using the main color image extraction algorithm to calculate the main color C(i, j) . For any two 4-connected image sub-block T(i, j) and

T(k ,l) (i - k =1 and j = l ; or j - l = 1 and i = k) ,

if its corresponds to the main color and in the HSV space to meet the following condition.

(1) Cj And Ci belong to the same color of magnitude, that is, its HSV components

hi = hj , si = s j , vi = v j ;

(2) Cj And Ci don't belong to the same color of magnitude, but satisfy

si * 3 + vi = s j * 3 + v j ,

and hi - hj = 1; or satisfy hi = hj , si = s j and vi , v j ∈ {0,1} .

We can say image sub-block T(i, j) and T(k ,l) are color connected. According to the concept of color connected regions, we can make each sub-block of the entire image into a unique color of connected set S = {Ri}(1 = i = M) in accordance with guidelines 4-connected. The set S corresponds to the color-connected region. For each color-connected region{Ri}(1 = i = M) , the color components R, G in RGB color space and H in HSV color space are respectively extracted the CCM at the direction d = 1;θ = 0 ,45 0 ,90 0 ,135 0 . The same operation is done with I (intensity of the image). The statistic features extracted from CCM are as

$$\text{Energy, E} = \sum_{i=1}^{p}\sum_{j=1}^{p} m(i,j)^2 \qquad (1)$$

$$\text{Contrast, I} = \sum_{i}^{p}\sum_{j}^{p}(i-j)^2 * \text{m(i, j)} \qquad (2)$$

$$\text{Entropy, S} = -\sum_{i}^{p}\sum_{j}^{p} m(i,j) * \log[m(i,j)] \qquad (3)$$

3.2.2 Texture feature extraction based on GLCM

GLCM creates a matrix with the directions and distances between pixels, and then extracts meaningful statistics from the matrix as texture features. GLCM texture features commonly used are shown in the following: GLCM is composed of the probability value, it is defined by P(i, j d, θ) which expresses the probability of the couple pixels at θ direction and d interval. When θ and d is determined, P(i, d,θ) is showed by Pi, j . Distinctly GLCM is a symmetry matrix; its level is determined by the image gray-level. Elements in the matrix are computed by the equation showed as follow:

$$P(i,j|d,\theta) = \frac{P(i,j|d,\theta)}{\sum_i\sum_j P(i,j|d,\theta)} \qquad (4)$$

GLCM expresses the texture feature according the correlation of the couple pixels gray-level at different positions. It quantificationally describes the texture feature. In this paper, four features is selected, include energy, contrast, entropy, inverse difference.

$$\text{Energy} \qquad E = \sum_x\sum_y p(x,y)^2 \qquad (5)$$

It is a gray-scale image texture measure of homogeneity changing, reflecting the distribution of image gray-scale uniformity of weight and texture.

$$\text{Contrast} \qquad I = \sum\sum (x-y)^2 p(x,y) \qquad (6)$$

Contrast is the main diagonal near the moment of inertia, which measure the value of the matrix is distributed and images of local changes in number, reflecting the image clarity and texture of shadow depth. Contrast is large means texture is deeper.

$$\text{Entropy} \qquad S = -\sum_x\sum_y p(x,y)\log p(x,y) \qquad (7)$$

Entropy measures image texture randomness, when the space co-occurrence matrix for all values is equal, it achieved the minimum value; on the other hand, if the value of co-occurrence matrix is very uneven, its value is greater. Therefore, the maximum entropy implied by the image gray distribution is random.

$$\text{Inverse difference} \qquad H = \sum_x\sum_y \frac{1}{1+(x-y)^2} p(x,y) \qquad (8)$$

It measures local changes in image texture number. Its value in large is illustrated that image texture between the different regions of the lack of change and partial very evenly. Here p(x, y) is the gray-level value at the coordinate (x,y)

International Journal of Scientific & Engineering Research Volume 4, Issue ł()ÜÕÌ-2013
ISSN 2229-5518

4 Two Tier Retrieval Approach

Our CBIR system operates at two tiers. In the first tier, the designed classifier categorizes the query image/images into one of the major disease types such as FL and NB. Once the disease category of the image is determined, the search for the query image can be carried out among the category relevant subtypes in the subsequent tier. For example, when the query image belongs to NB disease, database images in the first tier will be filtered according to the NB disease category. Then the subsequent search will be only performed on the NB category subset to retrieve the images from the correct category of the query images. In the second tier, we will use our proposed multi-image query and retrieval methodology to retrieve the images from the reference database in the order of their image-level visual similarities by preserving the slide-level semantic similarity.

4.1 First Tier: Classification of Disease Type With recursive SVM

An Recursive SVM-type classifier was employed to categorize the query image into one of the major disease type such as NB or FL using the extracted features, which are explained in Section III-A. Recursive SVM classifiers are well founded in statistical learning theory and have been successfully used for various classification tasks in computer vision. Their purpose is to find a decision hyper plane for a binary classification problem by maximizing the margin, which is the distance between the hyper plane and the closest data points of each class in the training set that are called support vectors. The hyper plane is chosen among all the possible hyper planes through a complex combinatorial problem optimization so that it maximizes the distance (called the margin) between each class and the hyper plane itself. As SVMs are restricted to binary classification, several strategies are developed to adapt them for multiclass classification problems [15] such as one-against-all classification and pair wise classification. In our recursive SVM classifier, we selected the radial basis function, which is one of the most frequently used kernels and it gives better results than other kernels for the categorization of our data. LibSVM MATLAB code [16] was used.

4.2 Second Tier: Retrieval using Decision Tree Algorithm

After determining the classes of query slides in the first tier, the next step is to retrieve the most relevant images from the database according to the main disease type of the query image set. Decision tree algorithm is a kind of data mining model to make induction learning algorithm based on examples. It is easy to extract display rule, has smaller computation amount, and could display important decision property and own higher classification precision. One benefit to decision trees is that the output is easy to explain to people without statistical training. Another benefit is that they allow you to look at interactions that occur in only some parts of the data. Each disease has higher level semantic annotations based on their histological grades such as Grade-I, Grade-II, and Grade-III in FL disease or D levels such as SR, UD, PD, and D in NB disease. Therefore, it is necessary to retrieve images related to their higher level semantic characteristics in order to provide more accurate results to the user of the CBIR system.

5 Dataset & Experimental Results

5.1 Dataset

Fig. 4 shows randomly selected sample images belonging to different histological grades of FL cases. The number of cropped images per slide is between 11 and 30 for FL cases and between 7 and 35 for NB cases. For FL slides, a team of experienced hematopathologists selected about 10 random microscopic high power fields (HPF) to interpret the disease grade in terms of the average number of centroblasts per HPF.

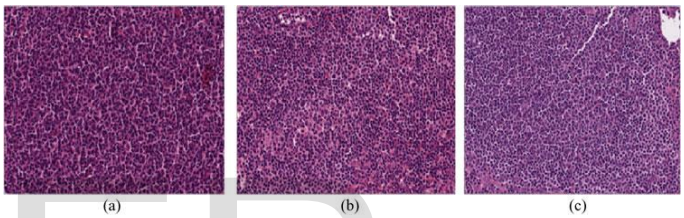

Fig. 2. Sample H&E-stained FL images associated with the three grades. (a) Grade I. (b) Grade II. (c) Grade III

For NB slides, pathologists pick the representative regions (images) from the whole slide and examine those regions at higher magnifications. The final decision for the differentiation grade of the entire slide is based on the grades of the sample images selected from that slide. Due to this differentiation grades, NB disease is differentiated to two subcategories such as SR and SP. SP subtype has three more subtypes such as D, PD, and UD.

5.2. Experimental Results

To evaluate the medical diagnosis systems, Sensitivity and Specificity criteria are usually used. The equations 7 and 8 define these two criteria [14]

$$Sensitivity = \frac{TP}{TP + FP} \quad (7)$$

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

In these equations, TP, TN, FN and FP show, the number of true positive, number of true negative, number of false negative and number of false positive samples respectively. In classification systems, accuracy is often used to evaluate the systems. Accuracy is the overall correctness of the system and is calculated as the sum of correct classifications divided by the total number of classifications as in Equation 9.

International Journal of Scientific & Engineering Research Volume 4, Issue ł()ÜÕÌ-2013
ISSN 2229-5518

$$Accuracy = \frac{items\ classified\ correctly}{all\ items\ classified} \quad (9)$$

TABLE-I

Average Classification Results (%)

Image type	Number of images	Sensi-tivity	Specificity	Accuracy
FL	50	95%	94.7%	96%
NB	50	93.5%	92.5%	98%

6 Conclusion

In this paper, we have presented a novel content-based microscopic image retrieval algorithm. This CBIR system can enable the user, e.g., a pathologist, to select multiple HPF regions from a suspected tissue and submit those images as a query to the CBIR system and retrieve the most relevant slides with their semantic annotations with higher accuracies. The results, achieved under those challenging conditions, are also promising for automatic and unsupervised selected query images based on their HPF regions. Application of the proposed weighting strategy, inspired by the IR theory, is not limited to microscopic images only, and can be also useful for any type of multiquery search and content-based retrieval systems.

References

[1] Hatice Cinar Akakin and Metin N. Gurcan "Content-Based Microscopic Image Retrieval System for Multi-Image Queries"IEEE transactions on information technology in biomedicine, vol. 16, no. 4, july 2012

[2] 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," Int. J. Med. Informat., vol. 73, no. 1,pp. 1–23, 2004.

[3] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, C. Roux, "Wavelet optimization for content-based image retrieval in medical databases," Proc. of ScienceDirect (Elsevier), Medical Image Analysis, Vol. 14, No. 2, pp. 227–241, 2010.

[4] S. Jai-Andaloussi, M. Lamard, G. Cazuguel, H. Tairi, M. Meknassi, B. Cochener, C. Roux, "content based medical image retrieval: use of Generalized Gaussian Density to model BEMD's IMF," World Congress on Medical Physics and Biomedical Engineering, Vol. 25, No. 4, pp. 1249-1252, 2009.

[5] K. W. Tobin, M. Abdelrahman, E. Chaum, V. Govindasamy, T. P. Karnowski, "A probabilistic framework for content-based diagnosis of retinal disease," Conference proc. of the IEEE, Engineering in Medicine and Biology Society, pp. 6744–6747, 2007.

[6] L. Zheng, A. Wetzel, J. Gilbertson, and M. Becich, "Design and analysis of a content-based pathology image retrieval system," IEEE Trans. Inf. Technol. Biomed., vol. 7, no. 4, pp. 249–255, Dec. 2003.

[7] L. Yang, O. Tuzel, W. Chen, P. Meer, G. Salaru, L. Goodell, and D. Foran, "Pathminer: Aweb-based tool for computer-assisted diagnostics in pathology," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 3, pp. 291–299, May 2009.

[8] S. A. Napel, C. F. Beaulieu, C. Rodriguez, J. Cui, J. Xu, A. Gupta, D. Korenblum, H. Greenspan, Y. Ma, and D. L. Rubin, "Automated retrieval of ct images of liver lesions on the basis of image similarity: Method and preliminary results.," Radiology, vol. 256, no. 1, pp. 243–252, 2010.

[9] M. M. Rahman, P. Bhattacharya, and B. C. Desai, "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback," IEEE Trans. Inf. Technol. Biomed., vol. 11, no. 1, pp. 58–69, Jan. 2007.

[10] C. Akgul, D. Rubin, S. Napel, C. Beaulieu, H. Greenspan, and B. Acar, "Content-based image retrieval in radiology: Current status and future directions," J. Digital Imag., vol. 24, no. 2, pp. 208–222, Apr. 2011.

[11] O. A. OConnor and J. M. Vose, "Indolent follicular lymphoma." Lymphoma Research Foundation, 2008.

[12] K. Belkacem-Boussaid, S. Samsi, G. Lozanski, and M. Gurcan, "Automatic detection of follicular regions in h&e images using iterative shape index," Comput. Med. Imag. Graph, vol. 35, pp. 592–602, 2011.

[13] H. Shimada, I. M. Ambros, L. P. Dehner, J.-I. Hata, V. V. Joshi, B. Roald,D. O. Stram, R. B. Gerbing, J. N. Lukens, K. K. Matthay, and R. P. Castleberry, "The international neuroblastoma pathology classification (the Shi-mada system)," Cancer, vol. 86, no. 2, pp. 364–372, 1999.

[14] H. Muller, N. Michoux, D. Bandon, A. Geissbuhler, "A Review of Content-Based Image Retrieval Systems in Medical Applications - Clinical Benefits and Future Directions," International Journal of Medical Informatics of the ScienceDirect (Elsevier), Vol. 73, No. 1,pp. 1–23, 2004.

[15] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multi-class support vector machines," IEEE Trans. Neural Netw., vol. 13, no. 2, pp. 415–425, Mar. 2002.

[16] C.-C. Chang and C.-J. Lin. (2001). LIBSVM: A library for support vector machines.[Online]. software available: http://www.csie.ntu.edu.tw/cjlin/libsvm.

[17] J. Naik, S. Doyle, A. Basavanhally, S. Ganesan, M. D. Feldman, J. E. Tomaszewski, and A. Madabhushi "A boosted distance metric: Application to content based image retrieval and classification of digitized histopathology," in SPIE Medical Imaging: Computer-Aided Diagnosis, 2009.

[18] A. CruzRoa, J. Caicedo, and F. Gonzlez, "Visual pattern analysis in histopathology images using bag of features," in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2009, pp. 521–528.

[19] N. Mehta, R. S. Alomari, and V. Chaudhary, "Content based sub-image retrieval system for high resolution pathology images using salient interest points," Int. Conf. Proc IEEE EMBS, vol. 1, pp. 3719–3722, 2009. [20] D. Iakovidis, N. Pelekis, E. Kotsifakos, I. Kopanakis, H. Karanikas, and Y. Theodoridis, "A pattern similarity scheme for medical image retrieval," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 4, pp. 442–450, Jul. 2009.

[20] H. Pourghassem and H. Ghassemian, "Content-based medical image classification using a new hierarchical merging scheme," Comput.Med.Imag.Graph., vol. 32, no. 8, pp. 651–661, 2008.