ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

An Improved Framework for Efficient Disease Prediction Using Content Based Image Retrieval

Halah Ozhakkal Latheef[1], Ambili K[2]

M.Tech Student, Dept. of CSE., CCET, Kerala Technical University, Kerala, India[1]

Assistant Professor, Dept. of CSE, CCET, Kerala Technical University, Kerala, India[2]

ABSTRACT: The framework for a system that automatically detects and classifies various diseases based on the images generated in the medical industry using significant and biologically interpretable difference in features is being proposed. The main objective of the proposed system is to automatically detect and classify various diseases based on the images that has been inputted to the system which are mainly scanning, X-ray kind of reports. The system is expected to predict what kind of a disease a person is more likely to be affected with based on his medical reports which are dominantly images. The image retrieval technique here is based on Content Based Image Retrieval (CBIR) using k-means clustering algorithm. The major steps include decomposition, feature extraction, clustering and prediction. Haar Wavelet Transform does the image decomposition and using mean and standard deviation calculations feature vectors are constructed. Once query image is uploaded, similar processes are repeated and from clusters formed, similarity calculations are performed. An additional feature to the system include Automatic Label Correction involving label correction in case of wrong uploads which otherwise could lead to wrong predictions and hence affect accuracy.

KEYWORDS: Content Based.Image Retrieval; Indexing; Automatic Relabeling;

I. INTRODUCTION

Prediction of Acute Diseases like Cancer, Heart Disease, Stroke, Haemorrhage etc., at its early stage has been a challenging task for medical practitioners all over for efficient diagnosis and treatment planning. Identifying these diseases manually varies based on expertise and various other factors. The percentage of people getting affected with diseases like cancer, problems with the functioning of heart has increased to much higher levels in the recent years and all these are mainly due to the change in lifestyles that took place in the recent years. The possibility of cure from these diseases has increased due to recent combined advancement in medicine and engineering. The chances of curing such chronic or acute diseases are primarily dependent on its detection and diagnosis. The selection of the treatment totally depends on the level of malignancy. Medical professionals use several techniques for detection of these diseases. These techniques may include various imaging techniques such as X-ray, Computer Tomography (CT) Scan, Positron Emission Tomography (PET), Ultrasound, and Magnetic Resonance Imaging (MRI).

Content-based Image Retrieval (CBIR) aims at describing the complex object information of digital images by non-textual features, which are applicable for efficient query processing. Recently, many approaches for content-based retrieval have been published, which are specially designed to support diagnosis in the medical field.

Development of digital technology has led to an increase in the number of images that can be stored in digital format. So searching and retrieving images from large image databases has also become more challenging. CBIR has gained increased attention from researchers in the recent years. CBIR is a system which uses visual features of images in large image databases and performs user's requests. Important features of images are colour, texture and shape which give detailed information about the image.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

II. RELATED WORK

In the past few years, there has been several literatures in the design of techniques for detecting cancer, haemorrhage, and such malignant diseases. Reddy and Satyaprasad [2] has focused on color and texture features for Content Based Image Retrieval. For improved performance of retrieval, they combined color planes with histogram. The main advantage was that it provided improved retrieval accuracy. Reshma Chaudary and Patil [3] suggested a framework that combined all the three shapes, color and texture features of image. Similarity computations were based on most similar highest priority principle.

Apostolos, Nikolaos, Aristidis and Andreas [4] used Relevance Feedback with SVM [5][6] and Feature combination to improve retrieval efficiency, as the number of feedbacks increases, the retrieval accuracy also increases. Nishant Singh, Shiv Ram Dubey, Pushkar, Jay Prakash [7] evaluated a two phase method for extraction of semantic information so as to overcome the limitations from Relevance feedback. The first phase involves creation of feature database of images and the second phase involves retrieval of images relevant to the query image.

Yildizer, Balci, Hassan and Alhajj [8] proposed efficient content-based image retrieval using Multiple Support Vector Machines Ensemble. The main aim here was to find good similarity between images. To find class probabilities it used SVM regression models. This technique was found suitable for handling large databases, and also reduced the dimensionality of feature sets.

In this paper, a framework that detects cancer, haemorrhage and heart related diseases are being detected based on the scan images being inputted. For decomposition and feature extraction, Haar Wavelet decomposition is used and for clustering k-means clustering is used. As an improvement for the accuracy of the system, the concept of Automatic Relabeling is used. This would enable the accurate entry of data into the training set. The images for training and evaluation of the system were collected from nearby hospitals and medical centres.

The sections following include the methods and modules involved in the system, and also brief description on the usefulness of the system in the medical field.

III. PROPOSED ALGORITHM

Early detection of diseases enables better planning of the treatment schedule. Early diagnosis and treatment increases the chance of cure. The proposed framework provides a system which enables medical practitioners to obtain a rough idea about the disease, that is, chances of occurrence of a particular disease so that treatment can be planned accordingly. The architecture of the system is as shown in the figure below (Fig.1.). Major modules in the system include: Indexing and Searching.

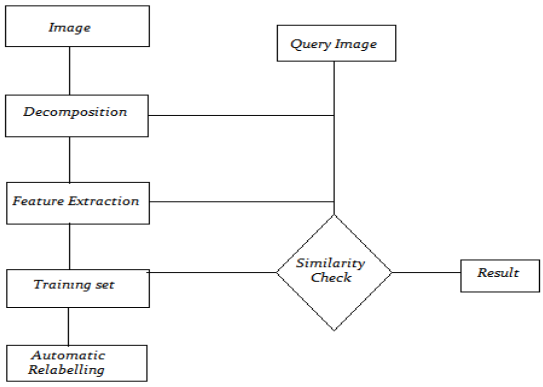

Fig.1. System Architecture

A. *INDEXING*

In the Indexing phase, both decomposition and feature extraction takes place. Each image in the database is represented as a set of image attributes i.e., RGB values of the pixels. The extracted features are stored as feature

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

vectors. Due to limitation in space and time, images are represented in reduced dimensions. Haar Wavelet Transform is used for calculating the feature vectors. Then k-means clustering algorithm is used for clustering the images based on their feature vectors considering the minimum Euclidean distance.

(i) Haar Wavelet Transform:

Haar Wavelet Transform is the simplest of Wavelet Transformations. The wavelets are useful in multi resolution analysis of the images because they are fast and give better compression where matrix represents a pixel in the image.

A Haar Wavelet transformation [9][11] decomposes an image into two components: average and difference. Each color in the image can be represented by considering the pixels as a point in space and from these matrices for each Red, Green and Blue components of RGB are constructed. This is then decomposed into four sub-matrices through row and column transformations.

The formula for calculating average at level 1 is given by-

$$a_n = \frac{+\quad +1}{\sqrt{2}}$$, where n=1,2,3…...n/2

And difference at the same level is given by-

$$d_n = \frac{-\quad +1}{\sqrt{2}}$$, where n=1,2,3 .....n/2

Once decomposition is done, then feature vectors can be constructed using mean and standard deviation of energy distribution (F-norm) of each sub-band at each level.

Given a square matrix A and A_i is the submatrix from it, then F-norm is given by-

$$\left| A_i \right|_F = [\ \Sigma_{=1} \quad \Sigma_{=1} \quad \left| \quad \right|^2 \ ]^{1/2}$$

Therefore, mean and standard deviation can be calculated as-

$$\text{Mean} = \Sigma \quad \frac{\quad}{\quad}$$

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma \quad (\quad - \quad)^2}{\quad}}$$

(ii) K-means Clustering Algorithm:

Clustering [10][11] partitions data points into clusters. Data points in a cluster are similar, whereas those in different clusters are different. It is a method used to automatically partition a dataset into k- groups. The basic step in k-means clustering is simple. Initially, determine the number of cluster K and assume the centroid or centre of these clusters. We can take any random objects as the initial centroids or the first K objects can also serve as the initial centroids. The algorithm repeatedly reassigns cases to clusters, so that the same case can move from cluster to cluster during the analysis.

Clustering is a very efficient and powerful tool to handle large data sets. It enables faster image retrieval and also allows search for the most relevant images in the large database. K-means is efficient in generating accurate results.

B. TRAINING

In the training step, the processes of indexing takes place, such that the properties of each image are stored in documents. Training mainly involves adding of images labelled with the disease name. An additional feature or enhancement to it involves Automatic Relabeling. It automatically pops up a dialogue box on wrong entry of a label for a particular image suggesting the more similar category of disease for that particular image.

Once images similar to the newly uploaded image are obtained, their similarity percentages are compared, and the one with the highest value is returned. If similarity is found to be greater than 75% then the image is uploaded with its corresponding label.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

C. SEARCHING

In the searching phase, when an image is uploaded, its feature vectors are computed. Then using the similarity criterion, these vectors are compared to the vectors of images in the training set. The most similar images are returned to a panel. The images returned are labelled with the corresponding disease and its percentage of occurrence. The classification of the most similar image would be returned as the class of the newly uploaded image.

IV. RESULTS AND DISCUSSION

The proposed methodologies were implemented using Net Beans IDE. It uses Haar Wavelet transformation for decomposing and feature extraction. Haar Wavelet is used due to its simplicity. It transforms an image into a matrix in which each element of the matrix represents a pixel in the image. The main advantage of using haar wavelet is its reduced computation time, which is improved speed for computation. Also, its simplicity and efficiency are also highlighting features.

After feature extractions, clustering of similar features are done. It is done using k-means clustering algorithm. It is simple and computationally faster. It is also easier to implement. They also minimize the Euclidean distance to the centroids of the cluster.

During the training step, a feature called automatic relabeling takes place, such that the wrongly uploaded image may be corrected automatically with user's consent by generating a dialogue box on wrong entry and suggesting a correct one for it. This would enhance the accuracy of the system, as the entries in the training set are checked for its correctness during each entry.

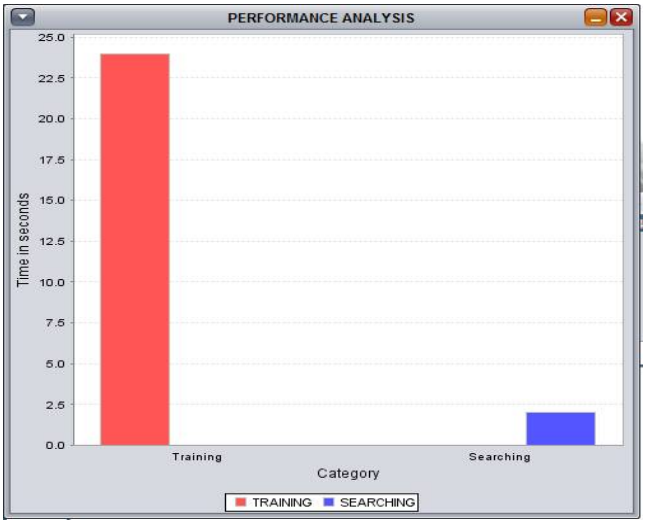

Fig.2. Performance

When a query image is uploaded, even it undergoes all these processes of decomposition, feature extraction and then calculate the Euclidean distance and then similarity check is performed. The most similar images to the one uploaded are returned together with the name of the disease and their percentage of occurrence.

A comparison graph showing time taken for training and searching has been plotted based on the real time values obtained. Modules against their time taken in seconds has been plotted to indicate their performance in the respective modules. Red indicates training module and Blue indicate searching module.

V. CONCLUSION AND FUTURE WORK

Compared to specific Applications, the proposed system provides a general CBIR technique for medical industry. Rather focussing on a single disease, the proposed system considers all diseases that can be assessed using the medical

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

images taken in hospitals for various purposes. A system of the proposed type would be very much acceptable as it could aid doctors in diagnosing, so that if detected early, better treatment plans can be formulated. Also, to improve the accuracy, automatic relabeling has been integrated, so that if any wrong uploads happen, it may be automatically being corrected. This in turn would improve the results provided by the system.

In the future, the proposed framework can be improvised to predict diseases other than those that can be predicted using images like diabetes, cholesterol etc.

REFERENCES

1. Rajesh Kumar, Rajeev Srivastava, and Subodh Srivastava," Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features", Journal of Medical Engineering, Vol. 2015, pp.14, 2015.
2. P.V.N Reddy, K.Satya Prasad, " Colour and Texture Features for Content Based Image Retrieval", International Journal on Computer Application and Technology, Vol.2, Issue 4, pp.1016-1020, 2011.
3. Reshma Chaudary, A.M Patil, " Colour and Texture Features for Content Based Image Retrieval", International Journal of Advanced Research in Electrical, Electronics And Instrumentation Engineering, Vol. 1, Issue 5, pp.386-392, 2012.
4. Apostolos Marakakis, Nikolaos Galatsanos, Aristidis Likas, and Andreas Stafylopatis, " Relevance Feedback for Content Based Image Retrieval using Support Vector Machine and Feature Selection", Springer, pp.942-952, 2009.
5. Sumiti Bansal and Er. Rishamjot Kaur, " A Review on Content Based Image Retrieval using SVM", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 7, pp.232-235, 2014.
6. K. Ashok Kumar and Y.V.Bhaskar Reddy, " Content Based Image Retrieval using SVM Algorithm", International Journal of Electrical and Electronics Engineering, Vol. 1, Issue 3, pp.38-41, 2012.
7. Nishant Singh, Shiv Ram Dubey, Pushkar Dixit, Jay Prakash Gupta," Semantic Image Retrieval by Combining Color, Texture and Shape Features", International Conference on Computing Sciences, pp.116-120, 2012.
8. Ella Yildizer, Ali Metin Balci, Mohammad Hassan, Reda Alhajj , "Content Based Image Retrieval using Multiple Support Vector Machine Ensemble", Journal on Expert Systems with Applications,pp.2385-2396, 2012.
9. Pasnur M.A, P. S. Malge, " Image Retrieval Using Modified Haar Wavelet Transform and K Means Clustering", Image Retrieval Using Modified Haar Wavelet Transform and K Means Clustering", International Journal of Emerging Technology and Advanced Engineering, vol.3, Issue 3, pp. 89-93, 2013.
10. Deepika Nagthane, " Content Based Image Retrieval system Using K-Means Clustering Technique", International Journal of Computer Applications & Information Technology, Vol. 3, Issue 1, pp.22-30, 2013.
11. Md. Jaffar Sadiq, Afshan Kaleem, Arif Hussain Mohammad, Mohammed Abdul Wajid, " Content Based Image Retrieval System using K-means and KNN approach by Feature Extraction", International Journal of Computer Science & Communication Networks,Vol 5, Issue 6, pp.391-399, 2015-2016.
12. Md. Iqbal Hasan Sarker, Md. Shahed Iqbal," Content-based Image Retrieval Using Haar Wavelet Transform and Color Moment", Smart Computing Review, vol. 3, Issue 3, pp.155-165, 2013.

BIOGRAPHY

Halah Ozhakkal Latheef is a student persuading final year in M.Tech Computer Science and Engineering, from Cochin College of Engineering and Technology under Kerala Technical University (KTU). She received Bachelor of Technology (B.Tech) degree in 2015 from MES College of Engineering under Calicut University, Kerala, India.

Ambili. K is an Assistant Professor in Computer Science and Engineering, Cochin College of Engineering and Technology, Kerala Technical University. She received Master in Technology (M.Tech) from MEA College of Engineering, Kerala, India. Her research interests are data mining, Big data etc.