Patch-Based Categorization and Retrieval of Medical Images

Zarina Sulthana and Supreethi K.P.

College of Engineering, Jawaharlal Nehru Technological University, Hyderabad,
Andhra Pradesh, India
{zarina.sulthana,supreethi.pujari}@gmail.com

Abstract. Utilization of the mining techniques in aiding the medical diagnosis by processing the medical images, Chest X-Rays in particular. The process involves - categorization of the images using patch-based visual words, collection into clusters using K-means clustering, and Image retrieval, involving comparison of the input image with the images in the dataset and retrieving the matched images along with the appropriate diagnosis associated with that particular medical case. If there is no data matching then this would be added to the existing database thus aiding the diagnosis of other such cases that might come up in the future. The retrieval may also be done based on Region-of-Interest (ROI).

Keywords: Mining, Medical Images, Categorization, Patch-based visual words, K-means clustering, Image Retrieval, Region-of-Interest.

1 Introduction

With the increasing influence of computer techniques on the medical industry, the production of digitized medical data is also increasing heavily. Though the size of the medical data repository is increasing heavily, it is not being utilized efficiently, apart from just being used once for the specific medical case diagnosis. In such cases, the time spent on the process of analyzing the data is also being utilized for that one case only. But if the time and data were to be utilized in solving multiple medical cases then the medical industry can benefit intensively from the medical experts' time in providing new and more effective ways of handling and inventing medical solutions for the future. This can be made possible by combining two most prominent fields in the field of computer science – *data mining techniques* and *image processing techniques.*

Medical imaging is the technique used to create images of the human body for medical procedures (i.e., to reveal, diagnose or examine disease) or for medical science. Medical imaging is often perceived to designate the set of techniques that noninvasively produce images of the internal aspect of the body. Due to increase in efficient medical imaging techniques there is an incredible increase in the number of medical images. These images if archived and maintained would aid the medical industry (doctors and radiologists) in ensuring efficient diagnosis.

V.V. Das and J. Stephen (Eds.): CNC 2012, LNICST 108, pp. 341–346, 2012.

The core of the medical data are the digital images, obtained after processing the x-ray medical images; these should be processed in-order to improve their texture and quality using image processing techniques and the data mining techniques may be applied in-order to retrieve the relevant and significant data from the existing million of tons of medical data.

2 Related Work

The diagnosis process is very time consuming and requires the expertise of a radiologist. Thus, automating the process of indexing and retrieval of medical images will be great boon to the medical community. The medical images are usually obtained in the form of X-rays using recording techniques, which add some unwanted extra data to the image like noise, air, etc., [2,5]. When these X-ray images are transformed into the digital format these disturbances also get converted into digital format and become a part of the image, which may adversely affect the process of generating the accurate data required when the images are processed for medical help. Thus these unwanted data needs to be separated from the images; and this can be done using *Image processing* techniques. Medical Image Mining includes the following phases: *pre-processing phase, bag-of-visual-words phase, Clustering phase and Retrieval phase.*

Image Digitization **-** images acquired as x-ray images need to be processed to remove the unwanted data from the images. The preprocessing phase is necessary in-order to get rid of the unwanted data from the image [4]. In the preprocessing phase, the image histogram equalization is used in-order to get the required clarity in the image [1]. Any medical image consists of: primitive features that are low-level features such as color, shape, and texture & logical features that are medium-level features describing the image by a collection of objects and their spatial relationships [8]. The images are divided into patches called visual words. The collection of all the patches is referred to as bag-of-words (visual words) [1, 10]. Features vector are created for the features of the image patches. These vectors are used in comparing the difference between the images; the difference is calculated in the form of Euclidean distance [9]. For different mining techniques, the results of feature extraction differs [3]. Based on the Euclidean distance, the images are segregated into multiple clusters. These clusters are utilized for retrieving a matching image during the retrieval phase.

3 Proposed Method

Preprocessing phase – includes the process of removing the unwanted data from the image and improve the quality of the images. This process of removing unwanted data (like stop-words in the data mining process) can be achieved by the techniques - *cropping, image enhancement and histogram equalization.* An X-ray image is obtained in the gray-scale. This gray-scale image is pre-processed using the *histogram equalization* in-order to improve its visual quality.

Bag-of-visual words phase – consists of the complete process of obtaining the *bag-of-visual words (patches) formation and feature extraction.*

Bag-of-Visual words (patches) formation – In this phase, the images are segmented into patches. Each segment will have more clarity when compared to the complete image. Each segment of the image is referred to as a patch (*visual word*). Each patch is a collection of the pixels in an image.

Feature Extraction - Post image enhancement and cropping, the images are obtained in a high-quality. For each obtained patch the hue, saturation, and brightness is calculated. Since it is a grey-scale image the hue and saturation counts remain zero. Thus ultimately the feature of brightness is considered for the future processing of the image. The patches are collected into a group and the average feature is calculated for the group. The group of the patches is formed by considering the adjacent patches obtained along the horizontal and vertical lines. Finally there would be certain number of groups of feature description of each image. These feature description groups along with the images are maintained in a database for the clustering process.

Clustering Phase - The images obtained from the previous phases are segregated into groups based on the similarity of the features extracted. Each segment thus formed is referred to a cluster. Initially a certain number of clusters are chosen, and then some randomly picked images are set as the centroid for each cluster. The images are compared against each other using the feature description groups and the Euclidean distance is calculated. The similar images are grouped together into a cluster. Having completed the process, the centroid of the cluster is recalculated based on the images present in it.

Retrieval phase - The medical images can be retrieved based on the feature comparison of the images stored in the clusters. Also image retrieval can be done for a specific region of interest (ROI) [5] using CAD (Computer-Aided Diagnosis) algorithms [1]. The medical image retrieval methods can be based on the type and nature of the features being considered. There are different categories of image retrieval methods such as, *Text Based Image Retrieval (TBIR), Content Based Image Retrieval (CBIR), and Semantic Based Image Retrieval (SBIR).*

Proposed Algorithm:

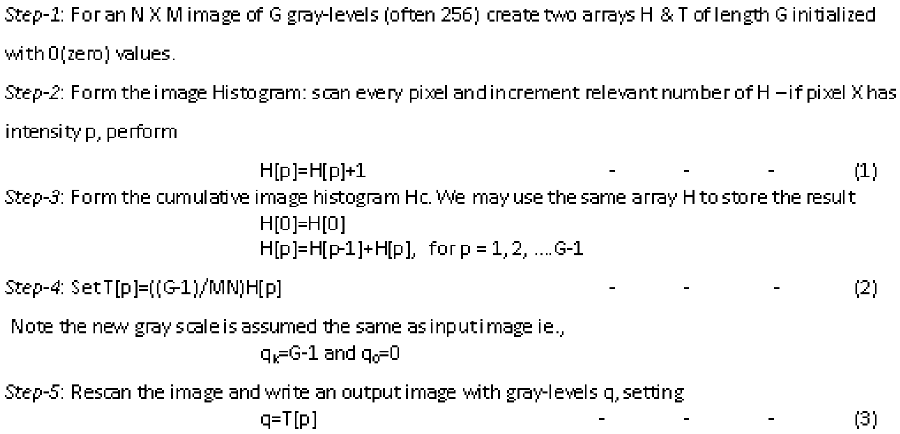
1. Divide the input image into patches
2. Apply histogram equalization to each patch of the image, in-order to improve the image quality.
3. Calculate the hue, saturation and brightness of the segments and generate the brightness vector for every patch of the image.
4. Store the images along with their vectors in the database.
5. Choose a random number of Clusters, and fill the clusters using K-means clustering.
6. For retrieving the matched images: repeat the steps 1 thru 5
7. Compare the image with the images in the clusters, in order to get the matching image
8. Retrieve the image and its recommended diagnosis from the database
9. Display the diagnosis report to the user.

4 Implementation

The work is implemented in the Java language using JDBC for the database connectivity and other features of Java pertaining to the images, and core Java.

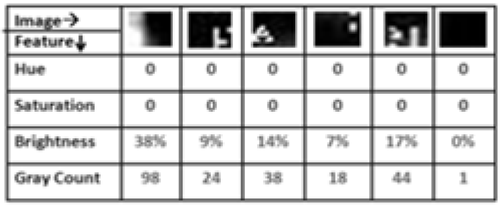
Pre-processing Phase: The first phase of the work includes, the pre-processing phase. The images which are to be preprocessed are collected in a directory as Buffered Images, and one by one those images are processed. Here for processing the histogram equalization technique is applied. The height and the width of the image is calculated and from this, the pixels are collected as *ints* (data type – integer) of the form 0xRRGGBB. Then from each pixel, the red, green, and blue components are extracted. From the data obtained the histogram is created. From the histogram equalization, we obtain much clear images. For performing histogram equalization the following algorithm is being used.

Histogram Equalization Algorithm:



```
Step-1: For an N X M image of G gray-levels (often 256) create two arrays H & T of length G initialized with 0(zero) values.
Step-2: Form the image Histogram: scan every pixel and increment relevant number of H –if pixel X has intensity p, perform
        $H[p]=H[p]+1$                                        -    -    -    (1)
Step-3: Form the cumulative image histogram Hc. We may use the same array H to store the result
        $H[0]=H[0]$
        $H[p]=H[p-1]+H[p]$,  for $p = 1, 2, ....G-1$
Step-4: Set $T[p]=((G-1)/MN)H[p]$                            -    -    -    (2)
 Note the new gray scale is assumed the same as input image ie.,
        $q_k=G-1$ and $q_0=0$
Step-5: Rescan the image and write an output image with gray-levels q, setting
        $q=T[p]$                                             -    -    -    (3)
```

Bag-of-Visual-words Phase: In this phase, the visual words (patches) are extracted from the equalized images, obtained from the previous phase. Initially, we obtain the width and height of the image, and then divide the image into patches based upon the Textons. Thus the required bag of visual words is formed.



Image→ Feature↓						
Hue	0	0	0	0	0	0
Saturation	0	0	0	0	0	0
Brightness	38%	9%	14%	7%	17%	0%
Gray Count	98	24	38	18	44	1

Feature Description

Feature description - For each patch in the image, the *HSB* (hue, saturation and brightness) are calculated. Since the hue and saturation count for the gray-scale images is zero, we take into account the *brightness feature*. The neighboring (along the vertical and horizontal line) patches are grouped into a *vector of patches*. For each patch vector we obtain the average brightness feature. This information obtained is stored into the database for the future calculations.

Cluster Phase: Based on the varied medical cases, we decide upon the number of clusters. Then randomly few images are picked up which are considered to be the centroid of each cluster. Then the similarity between the images is found out by calculating the *Euclidean distance* between the vectors of the images being considered. The images are said to be similar if their Euclidean distance is equal to the *threshold* value (assumed). Thus the similar images are collected into a cluster (here, the cluster is a directory). Having placed all the images into their respective clusters, for each cluster based upon the containing images, new centroid image is found.

Retrieval Phase – during the retrieval phase the query image is compared with the centroid images of all the clusters; and based on the closest matching centroid the respective cluster is picked up and the query image is compared with the containing images of the selected cluster, and the matching images is retrieved. Based on the retrieved image, its image id is found and the respective diagnosis report is retrieved from the database.

5 Conclusions

The process of diagnosis is a time taking process; and treatment is recommended only in the case where the patient is found to be affected; in such scenario, the time spent by the human expert in the diagnosis process is being wasted; instead if the diagnosis report generation process is automated, then the human expert time and experience may be utilized in a better way for the improvement of the medical field by inventing new effective ways of dealing with the diseases. Certain challenges are also faced due to the fact that dealing with the images is a very time consuming process; as well the storage requirement is very high. The work can be expanded in the future by taking into account various other features of the images like texture, shape etc.

References

1. Avni, U., Greenspan, H., Konen, E., Sharon, M., Goldberger, J.: X-ray Categorization and Retrieval on the Organ and Pathology Level, Using Patch-Based Visual Words. IEEE Transactions on Medical Imaging 30(3) (March 2011)
2. Bhadoria, S., Dethe, C.G.: Study of Medical Image Retrieval System. In: International Conference on Data Storage and Data Engineering. IEEE Computer Society (2010)
3. Fu, L.-D., Zhang, Y.-F.: Medical Image Retrieval and Classification Based on Morphological Shape Feature. In: Third International Conference on Intelligent Networks and Intelligent Systems (2010)
4. Antonie, M.-L., Zaïane, O.R., Coman, A.: Application of Data Mining Techniques for Medical Image Classification. In: Proceedings of the Second International Workshop on

Multimedia Data Mining (MDM/KDD 2001), in Conjunction with ACM SIGKDD Conference, San Francisco, USA (August 2001)

5. Unay, D., Ekin, A., Jasinschi, R.S.: Local Structure-Based Region-of-Interest Retrieval in Brain MR Images. IEEE Transactions on Information Technology in Biomedicine 14(4) (July 2010)
6. Suetens, P.: Fundamentals of Medical Imaging, 2nd edn. (2009)
7. Birkfellner, W.: Applied Medical Image Processing: A Basic Course
8. Jin, L., Hong, L., Lianzhi, T.: A Mapping Modeling of Visual Feature and Knowledge Representation Approach for Medical Image Retrieval. In: ICMA International Conference (2009)
9. Greenspan, H., Goldberger, J.: A continuous probabilistic framework for image matching (2001)
10. Avni, U., Greenspan, H., Goldberger, J.: Dense Simple Features for Fast and Accurate Medical X-Ray Annotation. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H., Tsikrika, T. (eds.) CLEF 2009. LNCS, vol. 6242, pp. 239–246. Springer, Heidelberg (2010)