International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2014): 3.05

A Content Based Image Information Retrieval for Medical MRI Brain Images based on Hadoop and Lucene (LIRe)

Abhay R. Palle[1], Dr. R. B. Kulkarni[2]

[1]Walchand Institute of Technology, Solapur, Maharashtra, India
[2]Walchand Institute of Technology, Solapur, Maharashtra, India

Abstract: ***CBIR technique is one of the important techniques for feature extraction in medical field in order to store, manage, and retrieve image data based on user query. Searching is done by means of matching the image features such as texture, shape, or different combinations of them. We propose a robust retrieval using a supervised classifier which concentrates on extracted features. Gray level co-occurrence matrix algorithm is implemented to extract the texture features from images. The feature optimization is done on the extracted features to select best features out of it to train the classifier. Hadoop platform is used to store large dataset of images. The purpose of this research work is to conclude normal and abnormal image using content based image retrieval on hadoop platform.***

Keywords: CBIR, Hadoop, Lucene

1. Introduction

1.1 Image Processing

Image processing operations can be roughly divided into three major categories, Image Compression, Image Enhancement and Restoration, and Measurement Extraction. It involves reducing the amount of memory needed to store a digital image. Once the image is in good condition, the Measurement Extraction operations can be used to obtain useful information from the image. Some examples of Image Enhancement and Measurement Extraction are given below. The examples shown all operate on 256 grey-scale images. This means that each pixel in the image is stored as a number between 0 to 255, where 0 represents a black pixel, 255 represents a white pixel and values in-between represent shades of grey. These operations can be extended to operate on color images.

1.2 MRI - Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a test that uses a magnetic field and pulses of radio wave energy to make pictures of organs and structures inside the body. In many cases MRI gives different information about structures in the body than can be seen with an X-ray, ultrasound, or computed tomography (CT) scan. MRI also may show problems that cannot be seen with other imaging methods.

1.3 Brain Images of MRI

Magnetic Resonance Images are used in detecting and tracking brain tumors. The tracking of the tumors is important especially when a patient is under medication in order to observe the changes that appear. Diagnostic brain MRI scans are usually performed by trained medical technologists who manually prescribe the position and orientation of a scanning volume. In this study, a fully automatic computer algorithm is described which compensates for variable patient positioning and acquires brain MRI scans in a predefined reference orientation. The human brain is among the most complex systems known to man. The rapidly growing technology in the past decade has made it possible for physicians to capture simultaneous responses from brain functions to test many long-standing brain theories. Almost all experiments in brain research have resulted in massive amounts of data. Often, neuroimaging and neurophysiological signals come in the form of large spatial and temporal data, also known as Multidimensional Time Series (MDTS). Very few studies in brain research and data mining have been tailored to exploit both spatial and temporal properties of these brain data.

1.4 CBIR

More and more attention is focused on Content-Based Image Retrieval (CBIR), which is a sub-problem of Content Based Retrieval (CBR). The tremendous growth of the numbers and sizes of digital image and video collections on Web is making it necessary to develop power tools for retrieving this unconstrained imagery.

In addition, CBIR is also the key technology for improving the interface between user and computer. CBIR systems have been developed in the recent years to organize and utilize the valuable image sources effectively and efficiently for diverse collections of images. Existing CBIR systems in biomedicine are designed to classify and retrieve images according to the anatomical categories of their content, i.e., head or chest X-ray images or abdominal CT images. CBIR for CT images of three types of liver lesions was investigated by incorporating semantic features observed by radiologist as well as features computationally extracted from the images.

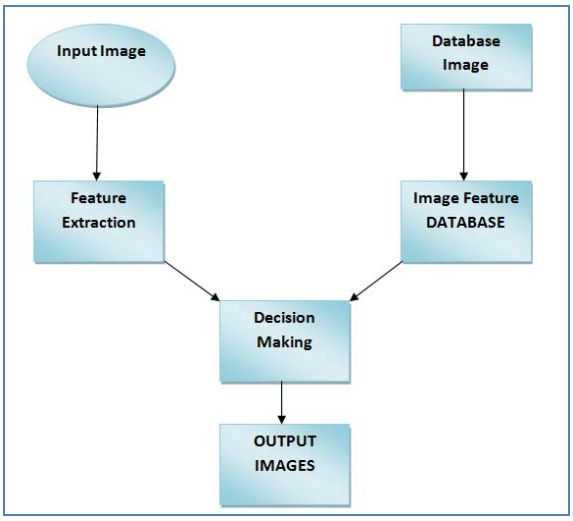

Figure 1: Content Based Image Retrieval

1.5 HADOOP

Hadoop consists of two components of HDFS and MapReduce, which are respectively the implement of Google's GFS and MapReduce. HDFS is a distributed file system for big data storage with outstanding scalability and fault tolerance. MapReduce is a distributed framework for data processing, especially big data. The MapReduce process consists of two steps, Map and Reduce. Splits of data are inputted into the Map process which will output intermediate key-value pairs. And then lists of these key- value pairs each of which has a common key are inputted into the Reduce process to output final key value pairs.

HDFS: HDFS is a java based distributed file system for big data storage that provides feature like scalability, self healing and fault tolerance.

A function called "Map," that concentrates distribution of the major task and creates map task for each range. Output of each map task is partitioned into key-value pair

A function called "Reduce," which collects the results and combines it to final form of the clusters

1.6 LUCENE

Lucene is a popular open source search library and often used to build up full-text search engine. Its powerful component of inverted indexing can also be used for constructing a CBIR system engine. Its powerful component of inverted indexing can also be used for constructing a CBIR system. In Lucene's indexing, Directory class stands for the directory storing index files; Document class stands for an indexing document, such as a text or a record of a table; Field class stands for different parts of the document and IndexWriter class is responsible for creating index for the document. In Lucene's indexing, Directory class stands for the directory storing index files; Document class stands for an indexing document, such as a text or a record of a table; Field class stands for different parts of the document and IndexWriter class is responsible for creating index for the document. In Lucene's indexing, Directory class stands for the directory storing index files; Document class stands for an indexing document, such as a text or a record of a table; Field class stands for different parts of the document and IndexWriter class is responsible for creating index for the document

1.7 OCTAVE / MATLAB

GNU Octave is a high-level interpreted language, primarily intended for numerical computations. It provides capabilities for the numerical solution of linear and nonlinear problems, and for performing other numerical experiments. It also provides extensive graphics capabilities for data visualization and manipulation. Octave is normally used through its interactive command line interface, but it can also be used to write non-interactive programs. The Octave language is quite similar to Matlab so that most programs are easily portable computations. It provides capabilities for the numerical solution of linear and nonlinear problems, and for performing other numerical experiments. It also provides extensive graphics capabilities for data visualization and manipulation. Octave is normally used through its interactive command line interface, but it can also be used to write non-interactive programs. The Octave language is quite similar to Matlab so that most programs are easily portable.

2. Literature Survey

Histological image retrieval based on semantic content analysis

The demand for automatic recognition and retrieval of medical images for screening, reference, and management is increasing. We present an intelligent content-based image retrieval system called I-Browse, which integrates both iconic and semantic content for histological image analysis. The I-Browse system combines low-level image processing technology with high-level semantic analysis of medical image content through different processing modules in the proposed system architecture. Similarity measures are proposed and their performance is evaluated. Furthermore, as a byproduct of semantic analysis, I-Browse allows textual annotations to be generated for unknown images. As an image browser, apart from retrieving images by image example, it also supports query by natural language.

Design and analysis of a content-based pathology image retrieval system

A prototype, content-based image retrieval system has been built employing client/server architecture to access supercomputing power from the physician's desktop. The system retrieves images and their associated annotations from a networked microscopic pathology image database based on content similarity to user supplied query images. Similarity 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. Current retrieval accuracy varies across pathological categories depending on the number of available training samples and the effectiveness of the feature set. The distance measure of the search algorithm was validated by agglomerative cluster analysis in light of the medical domain knowledge. Results show a correlation between pathological significance and the

image document distance value generated by the computer algorithm. This correlation agrees with observed visual similarity. This validation method has an advantage over traditional statistical evaluation methods when sample size is small and where domain knowledge is important. A multi-dimensional scaling analysis shows a low dimensionality nature of the embedded space for the current test set.

Content-Based Image Retrieval Using Multi resolution Color and Texture Features

In this paper, a content-based image retrieval method based on an efficient combination of multi resolution color and texture features. As its color features, color autocorrelo-grams of the hue and saturation component images in HSV color space are used. As its texture features, BDIP and BVLC moments of the value component image are adopted. The color and texture features are extracted in multi resolution wavelet domain and combined. The dimension of the combined feature vector is determined at a point where the retrieval accuracy becomes saturated. Experimental results show that the proposed method yields higher retrieval accuracy than some conventional methods even though its feature vector dimension is not higher than those of the latter for six test DBs. Especially, it demonstrates more excellent retrieval accuracy for queries and target images of various resolutions. In addition, the proposed method almost always shows performance gain in precision versus recall and in ANMRR over the other methods.

3. Methodology

We propose a robust retrieval using a supervised classifier which concentrates on extracted features.
Gray level cooccurance matrix algorithm is implemented to extract the texture features from images.
The classification is performed on the dataset and it is classified into three categories such as normal, benign and malignant.
To locate the abnormality and to reduce the training time of SVM classifier we insert clustering mechanism here.
The query image is classified by the classifier to a particular class and the relevant images are retrieved from the database.

3.1 Extraction of Texture Feature

A gray level co-occurrence matrix (GLCM) contains information about the positions of pixels having similar gray level values. A co-occurrence matrix is a two-dimensional array, P, in which both the rows and the columns represent a set of possible image values. A GLCM Pd[i,j] is defined by first specifying a displacement vector d=(dx,dy) and counting all pairs of pixels separated by d having gray levels i and j.

The GLCM is defined by:

- where nij is the number of occurrences of the pixel values (i,j) lying at distance d in the image.
- The co-occurrence matrix Pd has dimension n× n, where n is the number of gray levels in the image.

3.2 Classification by multi SVM (Support Vector Machine)

SVM maps input vectors to a higher dimensional vector space where an optimal hyper plane is constructed. Among the many hyper planes available, there is only one hyper plane that maximizes the distance between itself and the nearest data vectors of each category. This hyper plane which maximizes the margin is called the optimal separating hyper plane and the margin is defined as the sum of distances of the hyper plane to the closest training vectors of each category.

3.3 Similarity Measurements and retrieval

After getting the relevant image ids from KNN search the corresponding database index will be computed by similarity feature matching. With the help of that database index values the relevant images are retrieved and displayed.

4. Conclusion

Content Based Image retrieval method is used in order to find out abnormal MRI brain image on hadoop platform. Hadoop is used to increase size of the dataset as well as accuracy by comparing with the more images. Lucene is the indexing technique used for decreasing time required for searching. Further we can improve this work by implementing different classification algorithm and clustering algorithm.

References

[1] Lei, Zhang, Lin Fuzong, and Zhang Bo. "A CBIR method based on color-spatial feature." TENCON 99. Proceedings of the IEEE Region 10 Conference. Vol. 1. IEEE, 1999.
[2] Nazari, Mina Rafi, and Emad Fatemizadeh. "A CBIR system for human brain magnetic resonance image indexing." International Journal of Computer Applications 7.14 (2010): 33-37.
[3] R. S. Choras, Content-based image retrieval - A survey, BioMetrics, Computer Security Systems and Artificial Intelligence Applications, pp. 31-44, 2006.
[4] Shin, H. J., Eom, D. H., and Kim, S. S., 2005. One-class support vector machines-an application in machine fault detection and classification, Science Direct, volum 48, issue 2. Pages 395-408
[5] Sengur, Abdulkadir. "An expert system based on principal component analysis, artificial immune system and fuzzy k-NN for diagnosis of valvular heart diseases." Computers in Biology and Medicine 38.3 (2008): 329-338
[6] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," Communications of the Acm, vol. 51, pp. 107-113, Jan 2008
[7] Lux, Mathias, and Savvas A. Chatzichristofis. "Lire: lucene image retrieval: an extensible java cbir library." Proceedings of the 16th ACM international conference on Multimedia. ACM, 2008.
[8] Saha, Sanjoy Kumar, Amit Kumar Das, and Bhabatosh Chanda. "Cbir using perception based texture and colour measures." Pattern Recognition, 2004. ICPR 2004.

Proceedings of the 17th International Conference on. Vol. 2. IEEE, 2004
[9] Borthakur, Dhruba. "HDFS architecture guide." HADOOP APACHE PROJECT http://hadoop. apache. org/common/docs/current/hdfs design. pdf (2008).
[10] Lin, Yuanqing, et al. "Large-scale image classification: fast feature extraction and svm training." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.
[11] El-Dahshan, El-Sayed Ahmed, Tamer Hosny, and Abdel-Badeeh M. Salem. "Hybrid intelligent techniques for MRI brain images classification." Digital Signal Processing 20.2 (2010): 433-441.
[12] Cocosco, Chris A., et al. "Brainweb: Online interface to a 3D MRI simulated brain database." NeuroImage. 1997.
[13] Zhang, Dengsheng, and Guojun Lu. "Evaluation of similarity measurement for image retrieval." Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on. Vol. 2. IEEE, 2003.
[14] Hsu, Chih-Wei, and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines." Neural Networks, IEEE Transactions on 13.2 (2002): 415-425.
[15] Smith, Joh R., and Shih-Fu Chang. "Automated binary texture feature sets for image retrieval." Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference on. Vol. 4. IEEE, 1996