I J C T A, 9(28) 2016, pp. 425-430
© International Science Press

A REVIEW OF VISUAL INFORMATION RETRIEVAL ON MASSIVE IMAGE DATA USING HADOOP

K. Rajakumar* and D. Sudheer**

Abstract: Visual information retrieval on multimedia data and the related field of big data analytics have become increasingly important in both the academic and the research communities over the past two decades. Due to social networks the amount of data (text, video, audio, images, etc.,) being uploaded into internet is rapidly increasing, with Facebook users uploading over 2.5 billion new photos every month. Content based retrieval (CBR) [1] techniques are improvised for processing the multimedia data. Low-level and High-level algorithms are using to extracting the features from images. To achieve more efficiency for image retrieval techniques on huge number of images, big data concepts are applying. Applying distributed computational tools and parallel processing approaches can design efficient and reliable query engines based on image content.

Keywords: CBIR, feature extraction, Big Data, HDFS, Map Reduce.

1. INTRODUCTION

Image retrieval is the process of retrieving images from huge set of images based on similarity matching. Earlier days the similarity between images are measured by captions or description entered by the users along with images. The description of image is also called as metadata of the image. It is difficult to maintain as well as process meta data for huge set of images. The traditional image processing techniques like concept based image retrieval are failed to process efficiently. Content based image retrieval techniques are vastly using to extracting the features from multimedia data sets.

The amount of image data is rapidly increasing due to growth of social networks and satellite images.

This exponential data growth is given some challenges to processing and store the data. Keeping the data in multiple racks and maintaining replication is makes high availability of data. Processing the data on multiple racks in a parallel manner is an efficient technique. These two are can achieve by using hadoop open source computational framework. Hadoop is having three basic modules: 1. Hadoop distributed file system (storing the data on blocks), 2. YARN (resource management), 3. MapReduce (parallel processing paradigm).

2. APPLICATIONS OF IMAGE RETRIEVAL

Content Based Image Retrieval (CBIR) is one of the branch of image processing. Querying the images from the image datasets based on the semantic match of features. Content based search engines are most efficient than the key based search engines. There are several applications for CBIR, 1. Medical Image Databases, Scientific Databases, General Image Collections for Licensing, The World Wide Web. Medical

* Associate Professor, School of Computer Science and Engineering, VIT University, Vellore.
Email: rajakumar.krishnan@vit.ac.in
** Research Scholar, School Of Computer Science and Engineering, VIT University, Vellore.
Email: devulapalli.sudheer2016@vitstudent.ac.in

image databases like CT scan, MRI, ultrasound, visible human databases can efficiently process by using CBIR algorithms. The CBIR is also useful in scientific databases like earth sciences, satellite picks geographical images and send to the database servers. Based on the required features images are retrieved from the databases. Google using the QBIC (Query by Image Content)

3. FEATURE EXTRACTION

Content based image retrieval will done based on the visual content. Visual content like color, shape, texture.Biomedicine, Military, Education, Web image classification and searching are some of the areas where the CBIR technique finds its prime importance**.**

Color: There are many approaches to retrieve the image based on the color. Color histogram is used to compute the propagation of pixels of each color with in the images. The color histogram for each image will stored in the database.

Texture: Texture is nothing but the visual patterns associated with the image. Texture is used to distinguish between the areas of image with same color. There are number of texture features such as degree of contrast, directionality and regularity, coarseness, directionality and randomness etc.

Shape: Shape is one of the low level features of the image which is used to measure the shape of specific object. The natural objects are generally recognized by their shape. There are different types of shape features such as circularity, convexity, Lake Factor, direction, eccentricity, relative size etc.

Feature Extraction: Feature extraction is the process of transformation of the input data into set of features. In feature extraction method, the color, texture and shape features of the image are extracted. MATLAB, SciLab, NumPy these are the software available for feature extraction.

Similarity matching: Similarity matching is the process of approximating the solution based on computation of similarity function between a pair of images. Euclidean distance is the method for similarity matching in which the distance between two points are calculated. According to the distance the images are retrieved. The images with the less distance are provided as an output.

4. CBIR ARCHITECTURE USING HAOOP

Big data is a term to define the exponential growth of data. The large amount of data from petabytes of data considered as big data. The large amounts of data is failed to process with our traditional database management systems because of ETL problem. For multimedia data it will be more difficult to process the large data sets using traditional tools. Hadoop is open source framework which can process the bigdata efficiently. Hadoop having three main modules: 1. HDFS, 2. MapReduce.

HDFS: hadoop distributed file system is storage module along cluster of nodes. Hdfs is block structured storage file system. individual files are broken into blocks of a fixed size. These blocks are stored across a cluster of one or more machines with data storage capacity. [2]

MapReduce: MapReduce is parallel programming paradigm on large amounts of data. Mapreduce is having two phases map phase and reduce phase. Map phase is responsible to split the input data into key/value pairs and map the data. How many blocks are there in a cluster for dataset that many map instances will create and execute parallel on cluster. Reduce phase will combine all the map outputs and then it generate an user required final output.[3]

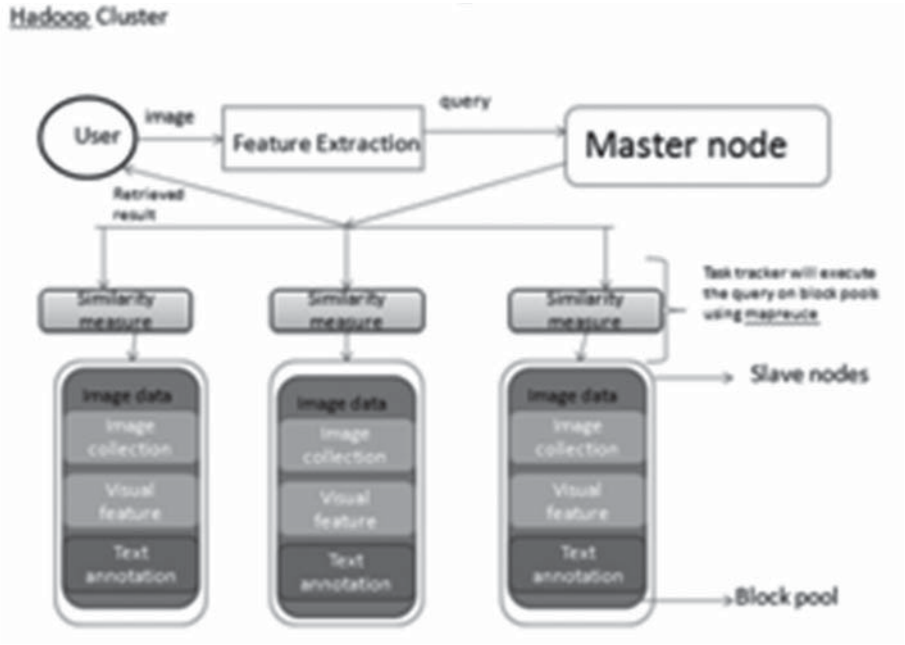

Figure 1. Image Retrieval System Using Hadoop.

Figure 1 illustrates that the haoop user gives a query image to feature extraction tool, later that results is given to master node of hadoop cluster. The master node will create instances of user query based on the image data blocks exists in a cluster. Suppose there are 3 slave nodes in a cluster. The large amounts of image data is segmented as blocks and stored in a three nodes then the master send the user query to all three slave nodes to process parallel.

5. LITERATURE REVIEW

Sayyed Mojtaba Benaei and Hossien Kardan Moghaddam illustrated that the amount of image data has grown considerably in recent years due to the growth of social networking. One of the well-known examples in this field is the generating PDF files from scanned daily archive of the New York Times in 2007. In this case 11 million photos with a volume of about 4 terabytes were converted to PDF only in 24 hours by using 100 nodes of Amazon Cloud Computing. The huge volume of visual data in recent years and their need for efficient and effective processing stimulate the use of distributed image processing frameworks in image processing area.[4]

Wichian Prem chaiswadi, AnuchaTung katsathan, and SarayutIntarasema says to deal with realistic situations, we present a joint querying and relevance feedback scheme based on the both high-level and low-level features of images for an on-line content-based image retrieval system. Hadoop mapreduce will reduce the processing time to retrieve the images from large datasets.[5]

Hinge Smita, Gaikwad Monika, Chincholkar Shraddha says Thousands of images are added to the image database and internet through the various digital devices. So we need an effective approach for storing and handling those images. Therefore we have described a novel approach of Content based Image Retrieval of large dataset by using the hadoop MapReduce for parallel processing. [6]

Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma illustrated that Research in content-based image retrieval (CBIR) in the past has been focused on image processing, low-level fea- ture extraction, etc. Extensive experiments on CBIR systems demonstrate that low-level image features cannot always describe high-level semantic concepts in the users' mind. It is believed that CBIR systems should provide maximum sup- port in bridging the 'semantic gap' between low-level visual features and the richness of human semantics.[7]

AnuchaTungkasthan, WichianPremchaiswadi taken histogram calculation to process the images on large clusters using hadoop mapreduce. They illustrates the distributed CBIR system processes two main functions using mapreduce framework to increase the performance of calculation. The applied feature extraction and similarity measurement using mapreduce.[8]

Said Jai-Andaloussi, AbdeljalilElabdouli, AbdelmajidChaffai, Nabil Madrane, Abderrahim Sekkaki says that medical imaging systems produce more and more digitized images in all medical fields. Most of these images are stored in image databases. CBIR system (Content-Based Image Retrieval) is one of the possible solutions to effectively manage image databases. Hadoop framework is used to store images and their features in column-oriented database HBase, and utilizes MapReduce computing model to improve the performance of image retrieval among massive image data. [9]

Andrey Sozykin, Timofei Epanchintsev illustrated the sizes of image collections are increasing dramatically and reaching petabytes of data. Such large volumes cannot be analyzed on personal computer within a reasonable time. Therefore, processing of modern image collections requires distributed computing. This paper presents a MapReduce Image Processing framework (MIPr), which provides the ability to use distributed computing for image processing. MIPr is based on MapReduce and its open source implementation Apache Hadoop. MIPr provides various forms of image representations in Hadoop internal format and the input/output tools for integration of image processing into Hadoop data workflow. The image formats in the MIPr framework are based on the popular image processing libraries. Furthermore, the MIPr includes the highlevel Image processing API for developers who are not familiar with Hadoop. This API allows to create sequential functions that process one image or a group of related images. The MIPr framework applies such functions to the large amount of images in parallel. In addition, MIPr includes MapReduce implementations of popular image processing algorithms, which can be used for distributed image processing without any software development. The MIPr framework significantly simplifies image processing in Hadoop distributed environment. [10]

Sridhar Vemula, Christopher Crick says that rapid growth of social media, the number of images being uploaded to the internet is exploding. Massive quantities of images are shared through multi-platform services such as Snapchat, Instagram, Facebook and WhatsApp; recent studies estimate that over 1.8 billion photos are uploaded every day. However, for the most part, applications that make use of this vast data have yet to emerge. Most current image processing applications, designed for small-scale, local computation, do not scale well to web-sized problems with their large requirements for computational resources and storage. The emergence of processing frameworks such as the Hadoop MapReduce platform addresses the problem of providing a system for computationally intensive data processing and distributed storage. However, to learn the technical complexities of developing useful applications using Hadoop requires a large investment of time and experience on the part of the developer. As such, the pool of researchers and programmers with the varied skills to develop applications that can use large sets of images has been limited. To address this we have developed the Hadoop Image Processing Framework, which provides a Hadoop-based library to support large-scale image processing. The main aim of the framework is to allow developers of image processing applications to leverage the Hadoop MapReduce framework without having to master its technical details and introduce an additional source of complexity and error into their programs.[11]

Venkat N. Gudivada, Vijay V. Raghavan says QBIC is a comprehensive, operational CBIR system based on the a priori feature extraction approach.The features are extracted semiautomatically. The mature QBIC system allows queries on large image and video data• bases based on example images, user-constructed sketches and drawings, color and texture patterns, and camera and object motion. The ideas presented here will be useful to practicing engineers interested in developing CBIR sys• tems. Since subjectivity and imprecision are usually associ• ated withspecifying and interpreting subjective attributes, the query processor should be designed to deal interac• tively with these problems at the query specification or processing time. The query interface may be designed to function as a knowledge-based system to guide users through the query-specification process and to facilitate user-relevance feedback and incremental query reformulation, User involvement in providing the relevance feed back should be at a conceptual level.[12]

Yihun Alemu, Jong-bin Koh, Muhammed Ikram, Dong-KyooKim addressed advancement of database technology that incorporates multimedia data, an open question that always rose in the technology is how to retrieve/search images in the multimedia databases. There are a huge number of research works focusing on the searching mechanisms in image databases for efficient retrieval and tried to give supplementary suggestions on the overall systems. The growing of digital medias (digital camera, digital video, digital TV, e-book, cell phones, etc.) gave rise to the revolution of very large multimedia databases, in which the need of efficient storage, organization and retrieval of multimedia contents came into question. Among the multimedia data, this survey paper focuses on the different methods (approaches) and their evaluation techniques used by many of recent research works on image retrieval system.[13]

Table 2.
Advantages Over Traditional Systems

Traditional work on CBIR	*CBIR with Hadoop*
Feasible for small amount of data sets	Feasible for large amount of data sets
Data stores in a single node.	Data stores in multiple nodes
Parallel processing is not possible	Parallel processing possible
Query processing is slow	Query processing is fast.

5. CONCLUSION

As per my literature review many authors illustrated that image retrieval systems are work more efficiently on distributed computational tools, and applying queries in a parallel processing paradigm. Hadoop frameworks provides a reliable and high available data storage as well as processing with mapreduce algorithms. So content based image retrieval systems on massive data sets are more efficient by using hadoop framework than traditional approaches.

References

1. Peter apers an martin kersten, “content based retrieval in multimedia databases based on feature models”, University of twente, Enshede the Netherlands.
2. D.Sudheer, A.Ramana Lakshmi, “Performance evaluation of Hadoop Distributed File System In pseudo distributed mode and fully distributed mode”, IJCSE, vol-3, issue-9, 2015.
3. http://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/Map Reduce Tutorial. html.
4. SayyedMojtabaBenaei and HossienKardanMoghaddam, “Hadoop and its Role in Modern image processing”, open journal of marine science, 2014, 239-245.
5. WichianPremchaiswadi, AnuchaTungkatsathan, and SarayutIntarasema, “Improving performance of content-based image retrieval schemes using hadoop mapreduce”, IEEE transaction, issue 7, 2013.
6. Hinge Smita, Gaikwad Monika, ChincholkarShraddha, “ Content based image retrieval using hadoop mapreduce ”, International Journal of computer science trends and technologies, volume 2, issue 6, Nov-2014.

7. Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma, “ A survey of content-based image retrieval with high-level semantics ”, Elsevier, pattern recognition society, 2006.

8. AnuchaTungkasthan, WichianPremchaiswadi, “A Parallel Processing Framework using MapReduce for Content-Based Image Retrieval ”, eleventh international conference on ICT, IEEE, 2013.

9. Said Jai-Andaloussi, AbdeljalilElabdouli, AbdelmajidChaffai, Nabil Madrane, AbderrahimSekkaki, “Medical Content Based Image Retrieval by Using the HADOOP Framework ”, IEEE conference publication,2013.

10. Andrey Sozykin, Timofei Epanchintsev, “MIPr – a Framework for Distributed Image Processing Using Hadoop”, IEEE conference paper 2015

11. Sridhar Vemula, Christopher Crick “Hadoop Image Processing Framework” IEEE conference paper, 2015.

12. Venkat N. Gudivada, Vijay V. Raghavan, “content based image retrieval systems”, IEEE, book chapter, 1995.

13. Yihun Alemu, Jong-bin Koh, Muhammed Ikram, Dong-KyooKim, “ Image retrieval multimedia database: A survey”, Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2009.