Survey Paper on Hadoop- using a Biometric Technique Iris Recognition

Umesh S. Soni
IT Department
Dr.D.Y.Patil Institute of
Engineering and Technology,
Ambi, Pune

Yogesh Wagh
IT Department
Dr.D.Y.Patil Institute of
Engineering and Technology,
Ambi, Pune

Silkesha Thigale
IT Department
Dr.D.Y.Patil Institute of
Engineering and Technology,
Ambi, Pune

ABSTRACT

Iris recognition is an automated method of biometric identification. Now in current world most of the iris comparison systems uses sequential and parallel execution but when iris dataset is large i.e. Big Data then in that case it has certain deficiencies like speed, complexity of dividing data, handling large data and robustness. So we are implementing Iris recognition process using an open source technology known as Hadoop. Hadoop technology is based on most popular programming model used to handle big data i.e. MapReduce framework. Hadoop provides specific library for handling large number of images: Hadoop Image Processing Interface (HIPI) and it can be used to implement the proposed system. Hadoop Distributed File System (HDFS) is used to handle large data sets, by breaking it into blocks and replicating blocks on various machines in cluster. Template comparison is done independently on different blocks of data by various machines in parallel. Map/Reduce programming model is used for processing large data sets. Map/Reduce process the data in <key, value> format. Iris database is stored in a <key, value> text format. HIPI is a library for Hadoop's MapReduce framework that provides an API for performing image processing tasks in a distributed computing environment.

Keywords

Hadoop, HDFS, MapReduce, Iris Recognition, HIPI, parallel image search

1. INTRODUCTION

Big data analysis is one of the highly researched areas today, with an aim of taking benefits of different computational resources. Using Hadoop environment performance of biometric matching can be increase. A biometric technique called iris recognition is used, as it is one of the strongest methods of authentication. Hadoop, is an open source distributed computing system, to develop our model. Hadoop performs Map/Reduce framework in Java. MapReduce make easy to process big amount of data and it uses Hadoop File System for handle the data. There are various techniques to uniquely identify an individual based upon physical and behavioural traits in biometrics field. Biometric technologies differ but all works in a similar manner, gather unique physiological and behavioural characteristics of a person, and store it into a database or comparing to found already stored templates in database. Iris recognition [2] is one of the strongest methods of biometric authentication. Iris recognition systems are gaining interest because it is stable over time. Iris recognition technology provides positive identification of an individual, at extremely high priority. Iris scan has been developing an identification/verification system capable of positively identifying and verifying the identity of individuals. It uses the unique patterns of the human iris, shows promise of overcoming previous shortcomings. When considering operations such as search and comparing iris pattern in iris images stored in a database, there are constraints on what can be done for improving the performance of single computers to make them able to process large-scale information. Therefore, the advantages distributed processing of an image database by using the computational resources of a distributed computing environment should be considered. Nowadays, for reasons such as ease of programming, by using the function MapReduce [5] on the Hadoop system, open-source cloud based systems that process data across multiple machines in a distributed environment have been studied for their application to various database operations. In fact, Hadoop [1] is in use all over the world. Studies using Hadoop have been performed to treat one as a text data file or multiple files as a single file unit, such as for the analysis of large volumes of DNA sequence data, converting the data of a large number of still images to PDF format, and carrying out feature selection/extraction in astronomy. Hadoop provides specific library named as Hadoop Image processing Interface for handling large number of small size images. In HIPI all can create image bundle of small size images and combine all images into one bundle and then divide that bundle in the cluster. Each data node will contain part of bundle and task tracker have to process only that much part of dataset. These examples demonstrate the usefulness of this system, which is due to its having the ability to run multiple processes in parallel for load balancing and task management.

2. HADOOP AT A GLANCE

2.1 Parallel and Distributed Processing on Hadoop

As the structure of the system, Hadoop [1] consists of two components, the Hadoop Distributed File System (HDFS) and MapReduce, performing distributed pro- cessing by single-master and multiple-slave servers. Map Reduce is divided into two elements, namely Job Tracker and Task Tracker, and HDFS has also two elements, namely Data Node and NameNode.There is also a mechanism that checks the metadata for Name Node.

2.1.1 Job Tracker

Job Tracker manages job scheduling and cluster resources, monitoring on separate components.

2.1.2 Task Tracker

In the cluster Task Tracker is a slave node daemon that accepts tasks and returns the results after executing tasks received by Job Tracker.

2.1.3 Name Node

An HDFS cluster consists of a single Name Node, a master server that sets the file system namespace and regulates access to files by clients. Name Node runs file system name space operations, such as closing, opening and renaming files, directories. It also determines mapping of blocks to a Data Nodes

2.1.4 DataNode

The cluster also has a number of Data Nodes, usually per one node in the cluster. Data Nodes sets the storage that is attached to nodes on which they run. DataNodes also perform block creation, replication and deletion in response to direction from NameNode

2.1.5 SecondaryNameNode

SecondaryNameNode is a helper to the primary NameNode. SecondaryNameNode is responsible for supporting periodic checkpoints of the HDFS metadata.

2.2 Hadoop Distributed File System (HDFS)

The HDFS[3] is a distributed file system designed to run on commodity hardware. It is inspired by Google System. HDFS is composed of NameNode and DataNode. HDFS stores each and every file as a sequence of blocks (currently 64 MB default) with all blocks in a file the same size except for the last block. Blocks containing a file are replicated for fault tolerance. The replication factor and block size are configurable for each file. Files stored in HDFS have only one writer at any given time.

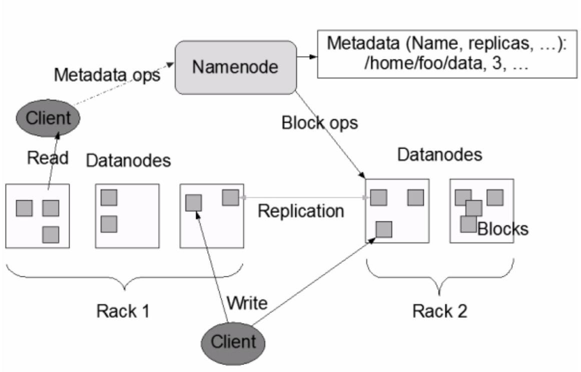


Figure 1: HDFS

2.3 MapReduce an Algorithm

MapReduce [4] (implemented on Hadoop) is a framework for parallel distributed processing large volumes of data. Using Map Reduce programming, it is possible to perform parallel distributed processing by writing programs involving the following three steps that are Map, Shuffle and Reduce. Because Map Reduce automatically performs inter-process communications between Map and Reduce process, and maintain balancing of the processes.

2.4 HIPI: Hadoop Image Processing Interface

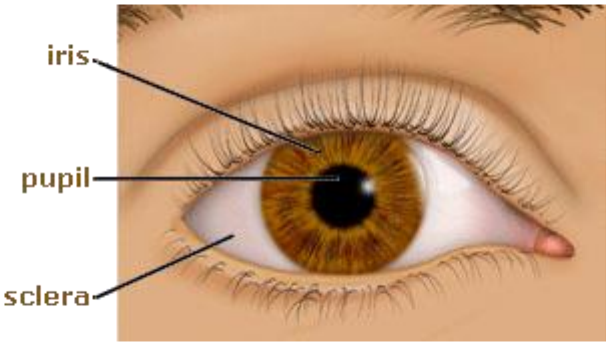
The amount of images being uploaded to the internet is rapidly increasing, with Facebook users loading over 2.5 billion new photos every month, however, applications that make use of this data are severely lacking[5]. Cur-rent computer vision applications use a small number of input images because of the difficulty is in acquiring computational resources and storage options for large amounts of data White et al. As such, development of vision applications that use a large set of images has been limited. The Hadoop MapReduce platform provides a system for large and computationally intensive distributed processing, though use of Hadoop's system is severely limited by the technical complexities of developing useful applications. To immediately address this, we propose an open-source Hadoop Image processing interface (HIPI)[5] that aims to create an interface for computer vision with MapReduce technology. HIPI abstracts the highly technical details of Hadoop's system and is flexible enough to implement many techniques in current computer vision literature. HIPI was created to empower researchers and present them with a capable tool that would enable research involving image processing and vision to be performed extremely easily. With the knowledge that HIPI would be used for researchers and as an educational tool, we designed HIPI with the following goals in mind. Hadoop uses a distributed file system to store files on various machines through-out the cluster. Hadoop allows files be accessed, however, without knowledge of where it is stored in the cluster, so that users can reference files the same way they would on a local machine and Hadoop will present the file accordingly.

3. "IRIS RECOGNITION" – A BIOMETRIC TECHNIQUE

A biometric system provides automatic recognition of an individual based on some sort of unique feature or characteristic possessed by the individual. Biometric systems are specially based on fingerprints, facial features, handwriting, the retina and iris.

3.1 The Human Iris

The iris[2] is a thin circular type, which lies between cornea and lens of the human eye. The iris is close to its centre by a circular aperture known as the pupil. The function iris is to control amount of light entering through the pupil, and it is done by sphincter and dilator muscles, which adjust size of pupil. The average diameter of iris is 12 mm, and the pupil size can vary from 10% to 80% of the iris diameter. Iris formation process begins during the third month of embryonic life. A unique pattern of the surface of the iris is formed during the year of life, and pigmentation of the stoma takes place for the few years.



3.2 Iris Recognition Process

An Iris recognition process consists following main sub processes:

1) **Image Acquisition:** - In this process eye image is captured. The image can be captured from live video camera or can be used already stored in memory or from a Dataset.

2) **Image Preprocessing:** -Preprocessing consists of image filtering and enhancement, iris image localization and normalization. Captured Image is converted into gray scale image if it is colored one. Canny edge detection algorithm is applied to detect the edge map of the image. For detecting inner and outer boundaries for pupil and iris, Hough Transform technique is used.

3) **Feature Extraction: -** Iris structure has complicated or complex and plentiful textures which can be extracted as features for coding. The Extracted Feature vector is compared with the already stored iris templates. Integer feature vector method can be used to compare with iris templates.

4. LITERATURE SURVEY REVIEW:

Weikuan Yu, Member, IEEE, Yandong Wang, and Xinyu Que, they proposed significant speeds up data movement from MapReduce and doubles the throughput of Hadoop. An algorithm called novel network-levitated merge algorithm is introduce to merge data without repetition and disk access [6].

I. Tomašić, A. Rashkovska and M. Depolli, "Using Hadoop Mapreduce in a Multicluster Environment" 2013. They proposed utilization of the MapReduce paradigm on a Hadoop installation extended across two clusters connected over the Internet [4]. Hadoop MapReduce has become one of the most popular tools for data processing.They measured execution times of MapReduce tasks in multi cluster environment, and compared them to the corresponding times obtained while only computers from a single cluster are used.

Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler 2010, They describe the architecture of HDFS and report on experience using HDFS to manage 25 petabytes of enterprise data at Yahoo! [3].The Hadoop Distributed File System (HDFS) is designed to store big data sets reliably, and stream that large data sets at high frequency bandwidth to user applications. In a big cluster, number of servers host directly attached storage and performs user application task. By providing computation and storage across multiple servers, the resource can grow in demand while remaining economical at every size.

Chris Sweeney Liu Sean Arietta Jason Lawrence "Hadoop Image Processing Interface" 2013, no of images being uploaded to the internet is rapidly increasing, with Twitter and Facebook users upload over 2.5 billion photos every month, but, applications make use of this data are lacking several [5].They propose an open-source Hadoop Image Processing Interface (HIPI) that aims to create an interface for computer vision with MapReduce technology.

5. PROPOSED SYSTEM

As all know that Hadoop works on record format, so complete database is stored in a format and uploaded in HDFS. We are going to take a big Data set of Iris images. For the result analysis we are going to setup two searching techniques for Iris Recognition application used on same data set. The first technique is a sequential execution, while other technique is Hadoop searching of the Iris Recognition application. We are going to test and compare both techniques on different nodes of different size. Time required for both type of searching techniques will be shown.

6. ACKNOWLEDGEMENT

The authors would like thanks to all the publishers and researchers for providing their resources available and teachers for their guidance also thank the college authority for providing the required infrastructure support.

7. CONCLUSION AND FUTURE WORK

Thus parallel search on Hadoop cluster takes less time than sequential search. Cluster with more nodes and higher configuration of nodes could be experimented on Hadoop with big data. While less number of nodes or dataset could reduce the performance of system. Experimentation on standard dataset i.e. big data to compare the performance on large size clusters. The experimentation with multicore nodes cluster and increase number of map and reduce task for fully utilizing multicore nodes. HIPI library can be used for solving other similar types of biometric recognition techniques and image processing problems.

1. The proposed system can be used to process large scale database of iris images.
2. Efficiency of the proposed work can be improved by adding more number of nodes and number of iris images.

8. REFERENCES

[1] Apache Hadoop. Http://hadoop.apache.org/

[2] Earnst, J. (n.d.) Iris Recognition Homepage. Retrieved February 15, 2005, from http://www.iris-recognition.org/

[3] IEEE Transactions on Parallel and Distributed Systems, Vol. 25, No. 3, March 2014, "Design and Evaluation of Network-Levitated Merge for Hadoop Acceleration, Weikuan Yu, Member,IEEE, Yandong Wang, and Xinyu Que.

[4] MIRPO 2013, May 20-24,2013 Opatija, Croatia "Using Hadoop MapReduce In A Multicluster Environment", I. Tomasic, A. Rashkovska and M. Depolli Jozef Stefan Institute/Department of Communications Systems, Ljubljana,Slovenia.

[5] **"**HIPI: A Hadoop Image Processing Interface for Image-based Map Reduce Tasks", Chris Sweeney Liu Liu Sean Arietta Jason Lawrence University of Virginia.

[6] The Hadoop Distributed File System, Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler Yahoo! Sunnyvale, California USA 2010.