International Journal of Pure and Applied Mathematics
Volume 118 No. 24 2018
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

A SURVEY ON IMAGE ANALYSIS USING MAPREDUCE AND HIPI

Mr. Omkar Sargar[1], Mr. Somnath Chenshetti[2],
Ms. Sapna Jagtap[3], Mr. Ajit Gund[4],
Jyoti Patil PhD[5],
[1,2,3,4]Research Scholar, Professor[5],
Department of Information Technology,
Jayawantrao Sawant College of Engineering,
Pune, India
omkar.sargar18@gmail.com,
somnaths9211@gmail.com,
sapnajag10@gmail.com,
ajitgund3199@gmail.com,
jyoti1199@gmail.com

May 27, 2018

Abstract

Due to increasing use of technology huge amount of data is generated per second, which may be structured or unstructured, termed as big data. Unstructured data mostly consists of images which cannot be processed using normal processing techniques. Well-known SIFT can be used for quick image processing. Extended Cluster Pruning algorithm can also be used for high dimensional indexing. Hadoops MapReduce is used for achieving parallelism in Image processing methods which provide highest fault tolerance as compared to other image processing methods. Hadoop image processing interface (HIPI) plans to make an interface for computer vision with MapReduce innovation. HIPI stores and processes images efficiently based on

Hadoop MapReduce platform. It provides a solution for how to store a large collection of images on the Hadoop Distributed File System (HDFS) and make them available for efficient distributed processing. This paper elaborates different image processing techniques which will help readers in choosing appropriate method in their development and research activity.

Key Words:Scale Invariant Feature Transform(SIFT), MapReduce, Hadoop image processing interface(HIPI), Extended cluster pruning(eCP), Hadoop.

1 INTRODUCTION

Nowadays everyone has a smartphone, increasing use of applications such as WhatsApp, Facebook, Instagram, etc. generates data in bulk. Mostly the generated data is unstructured i.e. Images or videos. Storing, analyzing and processing of such data is difficult using normal processing techniques. For quick processing image should be automatically analyzed by comparing it with the stored and indexed database. HIPI reposits and processes pictures with capability based on Hadoop MapReduce Platform. The answer for "In what way, Can huge amount of images be stored on HDFS" is given in it. And make them accessible for effective distributed processing. This paper describes different image processing methods like Scale Invariant Feature Transform (SIFT) and Extended cluster pruning algorithm which can be processed using HIPI based on Hadoop's MapReduce. Further sections will give more idea on each of technique.

2 IMAGE PROCESSING METHODS

For matching, analyzing and retrieving images of various image processing techniques we are focusing on methods such as SIFT and ECP.

SIFT-SCALE INVARIANT FEATURE TRANSFORM

In 1999, David Lowe at University of British Columbia came up with a new algorithm, Scale Invariant Feature Transform (SIFT) [2]. SIFT is a computer vision related algorithm to image analysis

and transformation. The name SIFT suggests that it is an invariant of scales which is used to transform the feature i.e. visuals. SIFT can recognize the object without changing the scale. It can also be used for mapping and navigation, gesture recognition, 3D mapping, etc. Key points play the key role to recognize the object in SIFT technique. SIFT includes 4 steps: A. Scale Spectra Extrema B. Locating Keypoint C. Orientation D. Keypoint Descriptor

A. Scale Spectra Extrema:

It identifies the focus which does not change to scale and angle utilizing Difference of Gaussians (DOG) which like wisely called as DOG Pyramid [1]. To construct the DOG pyramid the info picture is laced recursively with a Gaussian kernel of $= 1.6$. The end laced picture is down-inspected in each picture heading by factor of 2, and the lacing procedure is rehashed one more time. ***B. Locating Keypoint***

Key points are located from the image with respect to its scale. These key points are also referred as Interest points. These key points are obtained from scale space extrema of DOG. Local extremas are detected as key points.

Figure 1 Keypoints

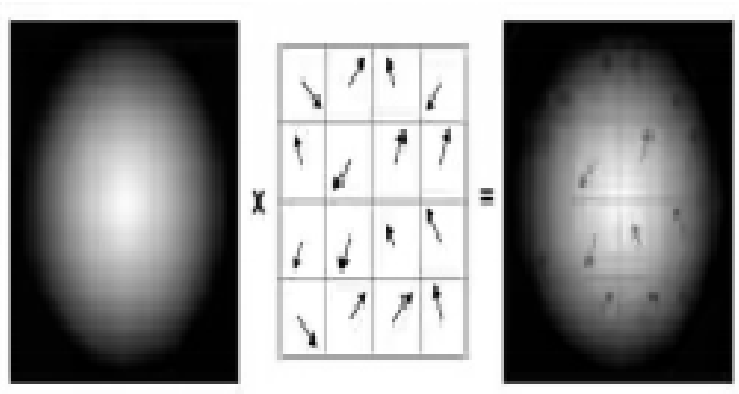

Figure 2 Orientation

C. Orientation

This section is all about rotation and different angles. Every point is prompted one or more angles based on regional picture gradient leanings. First, the Gaussian-flatten picture L(a, b, σ) at the key point's scale is grasped. For an image sample L(a, b) at scale σ, the gradient magnitude, m(a, b), and orientation θ(a, b), are figured before utilizing pixel differences:

$$\text{m(a,b)}=\sqrt{(L(a+1,b)-L(a-1,b))^2+(L(a,b+1)-L(a,b-1))^2}$$

$$\theta\text{(a,b)} = tan^{-1}\ \text{(L(a+1, b)-L(a-1, b), (L(a, b+1)-L(a, b-1)))}$$

The extent and course computations for the gradient are improved the situation each pixel in neighboring area around the key point in the Gaussian-blurred picture L [3]. When the histogram gets full, the orientations indistinguishable to toppest crest and local crests that are inside 80% of the most trustworthy crests are relegated to the key point.

D. Key-point Descriptor

To compute a descriptor vector for each key point with end goal for that the descriptor is exceedingly particular and incompletely

invariant to the rest of the varieties, for example, illumination, 3D viewpoint and so on. This progression is performed on the picture nearest in scale to the key-point's scale [3].

3 EXTENDED CLUSTER PRUNING(eCP) ALGORITHM

eCP is a high dimensional ordering methodology and it is especially identified with k-means approach[4]. Extended cluster pruning algorithm is partitioned into two stages: 1) Pre-processing of the information 2) query preparing. There are n number of point collection and in the pre-processing phase points are chosen randomly which are known as representatives. As showed in fig1.the representatives are denoted by dots as r1, r2, r3,, rn. What's more, the information focuses from the accumulation are then joined to the closest illustrative. From these connections we found the parcel of the information focuses into n clusters and there is one representative for each and every cluster. After this a leader is chosen from the representatives for each cluster which drives the cluster. Leaders are indicated by l1, l2, l3,, ln.

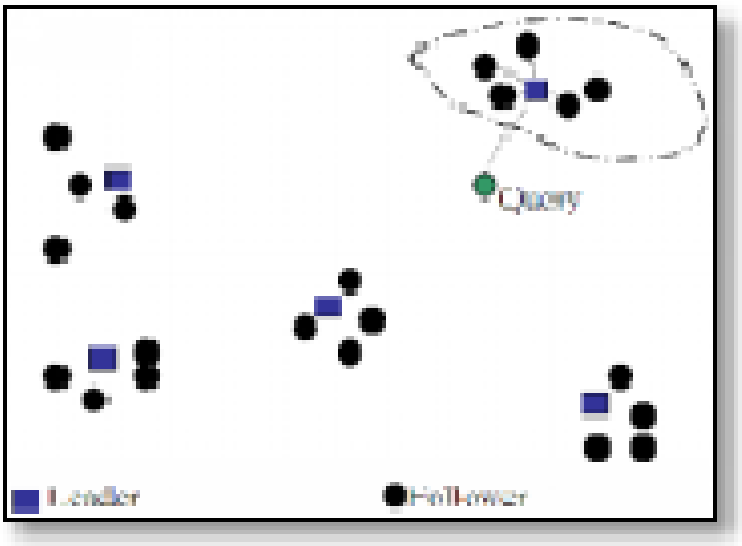

Figure 3 Leaders and followers (representatives)

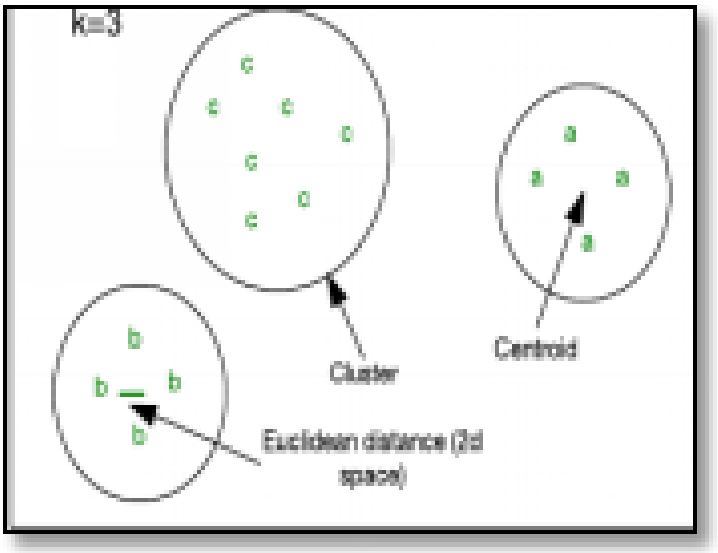

Figure 4 Cluster representation

To choose a leader there are three different ways such as, 1)A leader as a representative 2)A leader as a centroid of a cluster 3)A leader can be point of a cluster that is closest to the centroid of that cluster[5]. In the query processing we find the leader l which is closest to the query by computing distance from the query to the leader. Then attempt to find the nearest neighbors among the data points in the cluster for the query. After this we dont find the distance from the query to the data points and prune them away. Instinctively the neighbors from the query lies the cluster of the closest leader. For example, as shown in fig-2 Consider C points of collection called as representatives of C clusters and the normal number of information focuses called as Target Size (Ts) and C can be calculated as number of points divide by target size [5]. eCP then organizes the c representatives in a hierarchical levels that is in a tree structure and then read the representatives one by one. Cross the tree of representatives and are assigned out to the closest cluster representative delegate at the base of that tree. All these C clusters and tree of representatives being put away on the disk. The tree of representative is little and can fit in memory. It is important to explore down the tree of representative followed by the way showed at each level which is closest to query point. After that relating cluster is gotten. eCP can investigate the clusters that are nearest to the inquiry point. To find the quality of approximate search for nearest neighbors by finding the distance between the two points which gives best results of retrieval that is Euclidean distance metric[5]. Euclidean distance metric calculates distance between query image and the image present in the database. In the event that x and y are the Cartesian directions x = (x1, x2, x3 xn) and y = (y1, y2, y3 yn).Then the separation gap between the two focuses is computed as,

$$\mathrm{d(x,y)} = \sqrt{\Sigma_{i=1}^{n}(xi - yi)^2}$$

From the above computation as our aim is to find images from the database so here xi is the query picture and yi is the picture present in the database. If the distance between the query picture and database picture is zero then the same picture will be retrieved and if closer the distance similar image will be retrieved otherwise no picture will be selected.

A. Index creation and searching

eCP is split up into two main stages ,In stage 1 index tree is created and sorting out in a memory tree. Stage 1 is computationally modest and no dispersion is required. In this way assembling the memory progression is done once on a single machine. The resulting progression is sent to the different nodes presented in the construction of the index. Each of the node will use this progression for assigning the subsets of the data collections to clusters. In stage 2 splitting the data accumulation into free parts are assigned out to distinct nodes which accelerate the procedure. During stage 2 vectors are assigned to clusters. Comparing to feature descriptors of a query picture with descriptors of pictures in collection to search and pictures with closest descriptors is similar to a query picture.

System model:

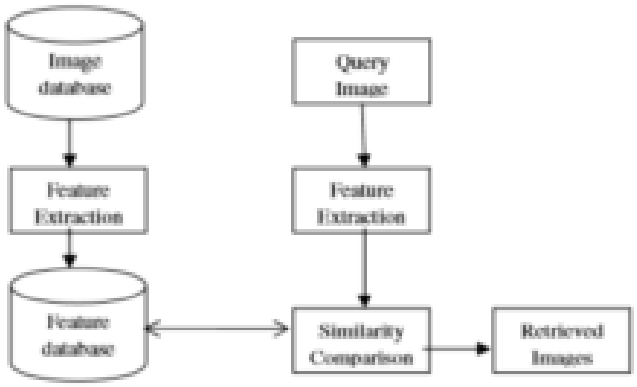

Figure 5 Image retrieval system

Algorithm:

1. Create a database consists of various pictures.
2. Extract the Texture and Intensity feature of each picture in the database.
3. Construct a combined feature vector for Texture and Intensity.

4. The new pictures formed are stored in another data-base called the Featured Databases.

5. Find the distance between feature vectors of query picture and that of Featured database pictures.

6. Sort the distance and Retrieve the top most similar pictures.

4 DISTRIBUTED PROCESSING METHODS

Large volume of data may be structured or unstructured is termed as Big data. It is information that is being created each second, so to process it, is the need of market. Unstructured information, for example, pictures is likewise produced at quicker pace. The MapReduce system performs huge volume of unstructured information on a Hadoop cluster. Hadoop is an open source structure which handles enormous information and it likewise gives high versatility, adaptation to internal failure, high accessibility and parallelism.

Map Reduce

To be specific 1) map phase 2) shuffle phase 3) reduce phase, Ma Reduce program is executed.

- **Map phase**

  The mapper's activity is to process the information. Basically information can be in a catalogue or in a file and is secured in the Hadoop Distributed file system (HDFS). The information record is passed to the mapper work line by line. The mapper forms the information and makes chunks of information.

- **Reduce phase**

  This phase is the mix of the Shuffle phase and the Reduce phase. To process the data that begins from the mapper is the reducers activity. In the wake of setting it up, passes on new course of action of yield, which will be passed in the HDFS.

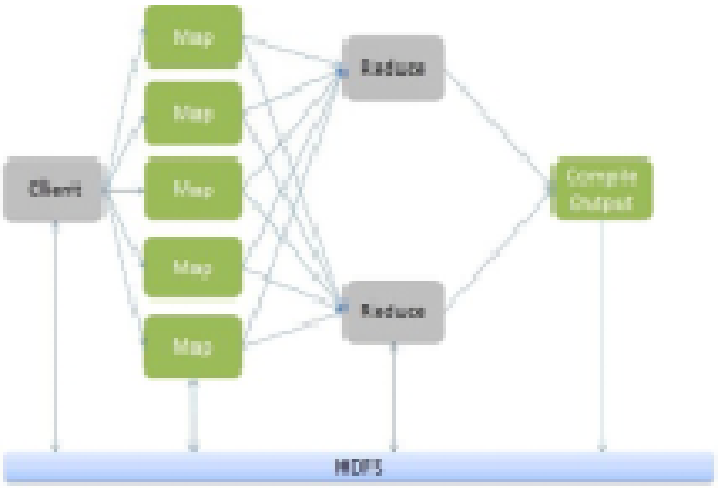

Figure 6 MapReduce architecture [6]

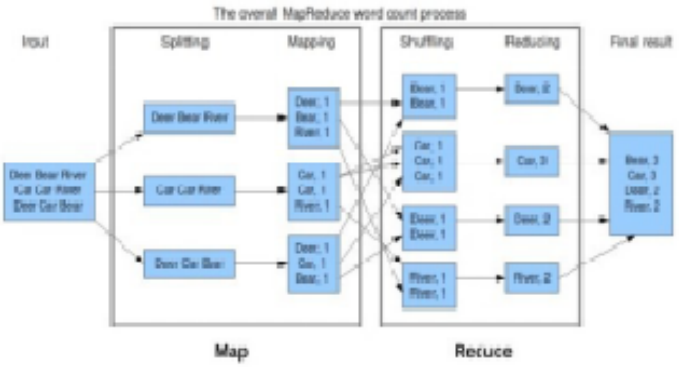

Figure 7 MapReduce Bear, deer, river and car example [6]

The above example quickly clarifies about MapReduce framework. For our thought how about we take a few expressions of a content record. We need to discover how often a word has showed

up in report. In the initial stage the input gets divided to separate the task among every one of the mappers as appeared in the figure. At that point each word gets distinguished and mapped to the main. The pair generated after the mapper task are known as tuples. At the first phase of mapper three words Deer, Bear and River are inputted. In this manner the yield of the mapper will be three key, value pairs with three unmistakable keys and value set to one. The mapping procedure continues as before in every one of the nodes. The output generated is then passed to the reducer. Then partition does the job of rearranging (shuffling) so all the tuples with unique key are passed to same node [7]. Reducer forms all the tuples to such an extent to the point that each one of the sets with same key are taken in tally and the consider gets refreshed the estimation of that particular key. In above case there are two sets with the key Bear which finally gets diminished to single tuple with the esteem equivalent to the check. All the yield tuples are then accumulated and written in the last yield record.

5 HADOOP IMAGE PROCESSING INTERFACE

HIPI is an image processing library intended to be utilized with the Apache Hadoop. MapReduce has a built-in parallel programming feature [8]. HIPI motivates efficient and high-throughput of image processing using MapReduce for parallel programs which are executed on a cluster. It gives an answer for how to store a substantial gathering of pictures on the HDFS and make them accessible for productive distributed processing [9]. HIB i.e. the HadoopImageBundle is the input given to HIPI. A HIB combines an already available set of images or from sources. A culling function works on the images to check whether they meet the specified criteria and eliminates those who dont adhere with [11]. The culling stage before the mapping stage gives the client a straightforward method to channel picture sets and control the sorts of pictures being utilized as a part of their MapReduce.

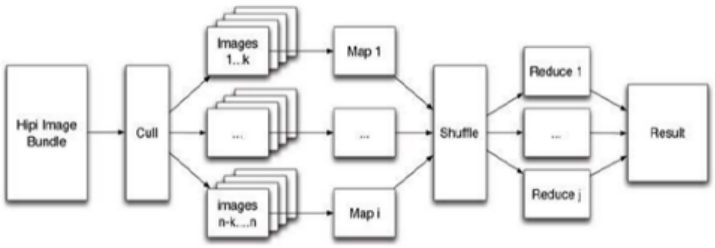

Figure 8 HIPI [10]

1. **Fig 8: HIPI [10] Hipi Image Bundle:** The input key to the HIPI framework is the HIB and it represents collection of images stored on the HDFS i.e. the Hadoop Distributed File System.

2. **Cull:** The initial stage of a Hadoop image processing interface program is a culling step, it filters the images in Hadoop based on criteria related to the big data.

3. **Images:** 3.The primary presentation for a collection of images on the Hadoop distributed file system. Map reduce is optimized to support efficient processing of large file system. HIB is actually compared two file stored on the system.

4. **Shuffle:** shuffle can start before and after map phase has finished to save same time. Reduce status is greater than 0 per cent but less than 3% when map status is not yet 100%.

5. **Mapper:** It takes the data and converts it into another set of data where each element is broken down into tuples.

6. **Reducer:** 6.Take the data from mapper and consolidate those data tuples into smaller set of tuples and gives the required output.

Following Tools are used in HIPI

1. hibImort

2. hibInfo

3. hibDump
4. hibDownload
5. hibToJpeg
6. covar

Analysis:
It is found that SIFT detects more number of features compared to other methods. It reduces duplication of calculation. Image analyzing and information extraction is improved with help of SIFT. The execution of eCP algorithm has been exhibited as far as speed picks up when we utilize more hardware. Hadoop as compared to other tools has a support for various types of input/output format support, compression techniques, and a customized scheduler. After studying HIPI tool we observed that it has powerful computing ability. HIPI gives efficient and accurate image search by MapReduce for large datasets. HIPI APIs also help in building applications efficiently.

6 CONCLUSION

We found that image analysis using map reduce which is already implemented in various projects can give best results using HIPI. We are going to use above studied methods in implementation of our project Image analysis using MapReduce and HIPI. This project can be implemented for crop disease detection, medical image analysis, satellite images, etc.

References

[1] Faraj Alhwarin, Chao Wang, Danijela Risti -Durrant, Axel Graser. Improved SIFT-Features Matching for Object Recognition. Institute of Automation, University of Bremen.

[2] Ebrahim Karami, Mohamed Shehata and Andrew Smith. Image Identification Using SIFT Algorithm: Performance Analysis against Different Image Deformations.

[3] https://en.wikipedia.org/wiki/Scale-invariant-feature-transform

[4] Archana Singh,Avantika Yadav,Ajay Rana. K-means with three different distance metrics. International Journal of Computer Applications, 2013.

[5] Diana Moise,Denis Shestakov,Gylfi Thor Gudmundsson,Laurent Amsaleg. Indexing and searching 100M images with mapreduce, 2013.

[6] https://www.dezyre.com/hadoop-tutorial/hadoop-mapreduce-tutorial-

[7] http://hadoop.apache.org/docs/r1.2.1/mapred-tutorial

[8] Jyoti S. Patil,Sunayana A. Mane. 3-D Image Analysis Using MapReduce. ICPC IEEE conference on pervasive computing, 2015.

[9] Dr. Peter Augustine, Christ University Banglore. Leveraging Big Data Analytics and Hadoop in Developing Indias Healthcare Services. International Journal of Computer Application, March 2014.

[10] Chris Sweeney,Liu Liu, Sean Arietta Jason Lawerence. University of Virgina, HIPI: Hadoop Image Processing Interface for Image-Based MapReduce Tasks.

[11] Lu Yu Zhe, Xu Chengcheng Huang, Prof. Ming-Hwa Wang. Image Search by Map Reduce. CEON 241 Cloud Computing Team Project, 2015.