Optimal Distributed Feature Extraction for Large Scale Images Using Hadoop Framework

Mrs.Suvarna L. Kattimani,Miss .SanaFalak I. **jakati ,** ***Miss.Heena M.*** **Sangtrash**[3]

[1]Assistant Professor Dept of CSE, BLDEA's Dr Halkatti College of Engineering & Technology, Vijaypur, Karnataka,(India)
[2]UG Scholar, DEPT of CSE,BLDEA's Dr Halkatti College of Engineering & Technology, Vijaypur, Karnataka,(India)
[3]UG Scholar, DEPT of CSE,BLDEA's Dr Halkatti College of Engineering & Technology, Vijaypur, Karnataka,(India)

Abstract

The main objective of this paper is to establish distributed feature extraction using large scale image database. In particular, the aim is to develop distributed feature extraction using state-of-the-art Bag-of-words features from a huge set of images. Feature extraction from images is key functionality required in many image/video processing and machine learning. Feature extraction generates an equivalent numerical representation of images. Due to availability of internet, mobile phones with camera, social networking sites huge amount of image and video data gets generated. Millions of images get generated within a very short span of time. This has led to research in the area of large scale image processing. If we want to process the million of images on high end servers it takes days together to characterize the images and even in some case the system will break. This has necessiated the need for distributed framework, which can be used for characterize huge image database in very short span of time. As part of results a table comprising of time taken by standalone system and cluster system will be generated. And show the behavior of hadoop to show how it relatively is better.

Keywords-Hadoop, MapReduce, Hdfs, Feature Extraction.

I.INTRODUCTION

Hadoop is an open source framework for writing and running distributed applications that process large amounts of data, the basic for writing a scalable, distributed data-intensive program Every day, in this modern era we’re surrounded by information in the form of data like people upload videos, take pictures on their cell phones, text friends, update their Facebook status, leave comments around the web, click on ads, and so forth. You may even be reading the book as digital data on your computer screen, and certainly your purchase of this book is recorded as data with some retailer. The exponential growth of data presents the challenges to cutting-edge business such as Google, Yahoo, Amazon, and Microsoft. They needed to go through terabytes and petabytes of data to figure out which websites were popular, what books were in demand, and what kinds of ads appealed to people. Existing tools were becoming inadequate to process such large data sets. Google was the first to publicize MapReduce a system they had used to scale their data processing needs. Computer, from the beginning of its invention, has been simplifying man’s life by performing many burdensome operations. Despite a very modest beginning, computer proved man’s friend and a servant of science, technology and industry. But as time progresses, even cutting edge technologies have to concede, get reduced to being obsolete in this case, the major reason being increased computational tasks. With necessity comes invention - to-day, we have access to faster and efficient computing technologies which have helped us to realize computational tasks never thought to be possible even a decade ago. One such application is image processing. Recently there has been a large number of papers published in the literature, which deals with extraction of local feature from images. These include SIFT[5] ,PCA-SIFT[6] ,colorSIFT[8,9] , SURF[7] and HOG[10] to name a few. But computer vision takes it a step further by applying image processing to various autonomous systems ranging from medical diagnostics to vigilance systems. Throughout the history of computer vision, amount of data to be processed versus available computational power has remained a major bottle-neck. The idea computer vision has gained steady ground today - thanks to faster computers. Feature Vector Extraction was done in Apache Hadoop using the new Map-Reduce paradigm. It was implemented by giving a higher level Hadoop container a Hadoop Sequence File [3] containing data of images to be processed as input to the program. These results were compared with Standalone Java Application that was run on a single machine.

II. LITERATURE SURVEY

In the internet service data base is most important part in that database image are stored in the world large data of images so that handling is very difficult just like ,the duplication of image it's increases the data size .We are all known about if that data was big the processing time also more With the proliferation of online photo storage and social media from websites such as Facebook and Picasa, the amount of image data available is larger than ever before and growing more rapidly every day . The billions of images available to us on the web. These images are improve, however, by the fact that users are supplying tags (of objects, faces, etc.), comments, titles and descriptions of this data for us. This information produces with an amazing amount of unprecedented context for images. The idea can be applied to a wider range of image features that allow us to examine and analyze images in a revolutionary way. The current processing of images goes through ordinary sequential ways to accomplish this job. The program loads image after image, The processing of data today is done by using oracle versions such as 9i,10G or by any another DBMS software. But with the increasing usage of internet all over the world the data on net is increasing rapidly. So, the processing of mass data is not possible by using any oracle software or any another existing DBMS software. the report generated after analysis will help the user to know about his usage. For analyzing the data & images we are going to use Hadoop technology.

III. EXISTING SYSTEM

Many data analysis software packages provide for feature extraction. Common numerical programming environments such as MATLAB, SciLab, NumPy provide some of the simpler feature extraction techniques but also pose many challenges like

1. Huge amount of cost is involved .

2. It is not available under free license.

3. Huge amount of time as the images increases.

4. Processing of huge amount is images is quite a impossible or is a risk to system stability.

IV. PROPOSED SYSTEM

Feature Vector Extraction was done in Apache Hadoop[1,2] using the new Map-Reduce paradigm. It was implemented by giving a higher level Hadoop container a Hadoop Sequence File containing data to images to be processed as input to the program. The images were used from Caltech101 (benchmark image database). These results were compared with Standalone Java Application that was run on a single machine.

The Apache hadoop cluster is setup on the machine and thus creates a pseudo cluster [4] on machine, once cluster is setup to programming using sift algorithm to extract features from images. Using stand alone machine extract the features from images. Store the images into Hdfs by converting images into sequence files, Sequence files is a flat file consisting of binary key/value pairs. It is extensively used in MapReduce as input/output formats. It is also worth noting that, internally, the temporary outputs of maps are stored using Sequence File [3,11].

The Sequence File provides a Writer, Reader and Sorter classes for writing, reading and sorting respectively.

Then MapReduce programming is carried out to extract the features using hadoop , MapReduce is a processing paradigm that builds upon these principles; it provides a series of transformations from a source to a result data set. In the simplest case, the input data is fed to the map function and the resultant temporary data to a reduce function. The developer only defines the data transformations; Hadoop's MapReduce job manages the process of how to apply these transformations to the data across the cluster in parallel. Though the underlying ideas may not be novel, a major strength of Hadoop is in how it has brought these principles together into an accessible and well-engineered platform [11].

V. SYSTEM ARCHITECTURE

The work break down structure for the project is as under:

1. Literature survey of image feature extraction, Bag-of-word model, distributed computing
2. Reading and understanding of Hadoop Framework
3. Establishing Pseudo cluster on single machine
4. Implementation of feature extraction using Bag-of-word model using Java
5. Selecting Set of category images from Caltech101 data (benchmark image database)
6. Testing the established feature extraction using Bag-of-Word using caltech images on standalone system and noting down the time taken for feature extraction.
7. Establishing distributed feature extraction using Hadoop
8. Testing the established feature extraction method on and noting down the timings.

VI. ALGORITHM

SIFT Descriptor: Algorithm can be divided into multiple parts.

1. Constructing a scale space:

This step generates several octaves of the original image. Each octave's image size is half the previous one. Within an octave, images are progressively blurred using the Gaussian Blur operator. Generally four octaves and five scales are taken.

2. Difference of Gaussian (DoG):

Two consecutive images in an octave are picked and one is subtracted from the other. Then the next consecutive pair is taken, and the process repeats. This is done for all octaves.

3. Locate Maxima/Minima in DoG images:

Maxima and minima are located in DoG images by making comparisons with adjacent pixels on the same scale and the scales above and below. This will result in at most 26 comparisons. Sub-pixel maxima and minima are obtained from the above points exploiting Taylor's expansion.

4. Eliminate Low Contrast Features and Edges:

To increase efficiency and robustness of the algorithm, low contrast points and edges are removed.

5. Assign Keypoint Orientation:

To assign an orientation we use a histogram and a small region around it. Using the histogram, the most prominent gradient orientation(s) are identified. If there is only one peak, it is assigned to the keypoint. If there are multiple peaks above the 80% mark, they are all converted into a new keypoint (with their respective orientations).

6. Generate SIFT Features:

A 16×16 window of in-between pixels is taken around the key point. Window is split into sixteen 4×4 windows. From each 4×4 window a histogram of 8 bins is generated. Each bin corresponding to 0◦ − 44◦ , 45◦ − 89◦ , etc. Gradient orientations from the 4 ×4 are put into these bins. This is done for all 4 × 4 blocks. Finally, the 128 values are normalized.

VII. FUTURE ENHANCEMENT

In the future, we might focus on images processing for large scale database and even for large scale satellite images obtained from social media like face book, Amazon etc. So our application will be beneficial in future to the following sectors:
1. Meteorological disaster - Violent, sudden and destructive change to the environment related to, produced by, or affecting the earth's atmosphere, especially the weather-forming processes.

2. Military navigation - study of traversing through unfamiliar terrain by foot or in a land vehicle.

3. Key point matching between two images – Match the key points against a database of that obtained from training images by finding the nearest neighbor i.e. a key point with minimum Euclidean distance.

4. Monitoring around the globe to extract discriminative information about regions of the globe for which GIS data is not available.

5. Feature vector generation and its study is very important aspect in machine learning [12].

VIII. REFERENCES

[1] http://lucene.apache.org/hadoop/2006

[2] http://en.wikipedia.org/wiki/Apache_Hadoop

[3] White, T.: Hadoop: the definitive guide, 2nd edn. O’Reilly Media, Sebastopol (2011)

[4]http://www.michael-noll.com/tutorials/
running-hadoop-on-ubuntu-linux-single-node-cluster/

[5] D. G. Lowe, ” Distinctive image feature from scale invariant keypoints ”, International Journal of computer vision, 2004,Vol. 60, 91-110.

[6] Yan Ke, Rahul Sukthankar, ” PCA-SIFT: A more distinctive representation for local image descriptors”, Proc. Conference Computer vision and pattern recognition, 2004,
511-517.

[7] Herbert Bay, Andreas Ess, TinneTuytelaars, LUC Van Gool, ” SURF: Speeded Up Robust Features ”, CVIU, Vol.110, No-3,2008,346-359.

[8] Evaluating Bag-Of-Visual-Words Representation in Scene Classification. Proc. International Workshop on Multimedia Information Retrieval, MIR07, 2007

[9] Gertjan J .Burghoutsa, Jan-Mark Geusebroek, ”Performance evaluation of local color invariants”, Computer Vision and Image Understanding, 113(2009) 48-62

[10] Navneet dalal and Bill triggs, ”Histogram of oriented gradients of human detection”, CVPR 2005.

[11] The Definitive Guide, Copyright © 2013 Packt Publishing Garry Turkington .

[12] Bishop, Christopher(2006). Pattern recognition and machine learning. Berlin: Springer.