



Retrieval of Satellite Images over Hadoop Distributed System

Vijay Mali, Akshay Jadhavar, Ganesh Mahimane, Diptee Chikmurge

BE Student, Dept. of Computer Engg, MIT AOE, Alandi, Pune University, Pune, India

BE Student, Dept. of Computer Engg, MIT AOE, Alandi, Pune University, Pune, India

BE Student, Dept. of Computer Engg, MIT AOE, Alandi, Pune University, Pune, India

Asst. Professor, Dept. of Computer Engg, MIT AOE, Alandi, Pune University, Pune, India

ABSTRACT: Within the last ten years we have been witnessing interest in research in the processing of Satellite Images. In today's digital era, integration of any form of data from heterogeneous datasets is considered as inevitable. This becomes complex when database is of large size. Considering the increased importance surface recognition, segmentation, as well as event detection in satellite images with a highly scalable system is becoming more and more desirable. In this project, we will construct a semantic taxonomy for the land-cover classification of satellite images. According to statistics it is estimated that 90% of the modern digital multimedia data has been created in the last 10 years. As an open source distributed computational framework, Hadoop helps in processing large amounts of images on an infinite set of computing nodes by providing necessary infrastructures, which gives advantages over the existing MATLAB system which is closed under the licence and is complex. The system will be divided into 2 parts i.e. Training part and Running part classifier. The Training part classifies subsequent satellite images such as Vegetation, Building, Pavement, water, snow etc. The images are further divided and classified and then stored onto hadoop nodes. Training part of classifier is implemented using MapReduce Framework of HADOOP and is based on Google Earth. The Running part of classifier can perform zoom-in, zoom-out as well as calculate the difference between old- new images.

KEYWORDS: Event detection, hadoop, MapReduce, segmentation, Training classifier, Running classifier

I. INTRODUCTION

We often wonder how google does its queries into their mountains of data, or how facebook is able to quickly deal with such a large quantities of Information. The answer takes us to the wild west of data management i.e. Big data. Many of us might or might not have heard about it, but one thing is sure that Big data, hadoop, mapreduce, but you can be sure that they can be a part of your regular conversations over the nest few years. This is because the 90% of the data was generated in the last 10 years. All the data in the world was mostly generated in the last 10 years and this accelerated trend is going to continue. All this new data is coming from smartphones, social networks, trading platforms or machines or other resources. In simple words, we are processing data parallel rather than in serial, so why is there need to call the wild west of data management? Over the long run, hadoop will be a part of our day to day information architecture. Hadoop is going to play a central role in Statistical analysis, ETL Processing and business intelligence. Hadoop is a software framework that allows you to not only store the large amount of data (>10TB) but also to efficiently access it using mapreduce. There are two main concepts when it comes to hadoop, 1. HDFS (Hadoop Distributed File System) 2. MapReduce.

The main advantages of this model are its simple programming structure, distributed file system, and distributed management which is failure resistant.

The events that occur on or beneath earth's surface are sometimes possible to detect. For example- The events such as flooding, tsunami and snow storm etc. They can be detected from the measurable changes in ground surface cover as a result of damage to existing structures. Satellite images take very huge amount of data storage. Large Scale Land-cover Recognition System Collects large amount of data of higher resolutions Satellite Images. It provides a collection of training data classifiers and performing subsequent image classification in distributed environment



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

This project aims to introduce an open source efficient way to retrieve, classify and search satellite images where it will use the concept of training and running classifier for a given area. The Training part classifies subsequent images such as Vegetation, Building, Pavement, water, snow etc. Large files are distributed and further divided among multiple data nodes. The Running classifier will perform zoom-in, zoom-out and calculates the difference between old-new images. Huge amount of similar images that can be found can be deleted over. The existing system MATLAB is closed under the licence in which a very large amount of cost is involved along with it being not providing user friendly environment and not being object oriented

We have used Google Earth for training classifier which involves features like zoom in zoom out boundary division, highlighting of buildings, rivers, roads and vegetations.

II. RELATED WORK

[1] LARGE SCALE SATELLITE IMAGE PROCESSING USING HADOOP DISTRIBUTED SYSTEM, Sarade Shrikant D., Ghule Nilkanth B., Disale Swapnil P., Sasane Sandip R., IJAR CET, ISSN : 2278 – 1323, Volume-3, Issue-3, March 2014.

This paper mentioned emphasis on using HDFS for retrieval of satellite images as well as setting up HDFS. It also has few contents about Image retrieval but no sufficient information about Image Processing techniques that can help in Image Segmentation and Feature Extraction.

[2] Banaei, S.M. and Moghaddam, H.K. (2014) Hadoop and Its Role in Modern Image Processing. Open Journal of Marine Science, 4, 239-245. <http://dx.doi.org/10.4236/ojms.2014.44022>

This paper gives a brief idea about the MapReduce which is a distributed data processing model which makes use of open source Hadoop framework for manipulating large volume of data. A very large amount of data in the modern world, especially multimedia data, implements new requirements for processing and storage as well. Hadoop allows the processing images on an infinite set of computing nodes with the help of necessary infrastructures.

[3] Object-Based Spatial Feature for Classification of Very High Resolution Remote Sensing Images, Penglin Zhang, Zhiyong Lv, and Wenzhong Shi., IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, ISSN : 1545-598X, Volume-10, Issue-6, March 2013.

It introduces a structural feature called object correlative index (OCI) which can be used for enhancing the classification of high resolution images. It uses spectral similarity to construct a useful OCI for describing the structural information objectively. If we Compare it with the generic features widely used in image classification, the classification approach based on the OCI spatial feature results in higher classification accuracy than those approaches that only consider spectral features or pixelwise spatial features, such as the pixel shape index and mathematical morphology profiles is observed. Experiments are conducted on QuickBird satellite image and aerial photo data, and results confirm that the proposed method is feasible and effective.

[4] Hadoop Map Reduce for remote Sensing Image Analysis, Mohamed H. Almeer, IJETAE, ISSN: 2250–259, Volume-2, Issue-4, April 2012.

University of Qatar tested and utilized Image processing algorithms related to remote sensing on the Hadoop MapReduce parallel platform by using an experimental 112-core high-performance cloud computing system that is situated in the Environmental Studies Centre. Although there is a considerable research in the utilization of the Hadoop platform for image processing rather than for its original purpose of text processing, it had never been proved that Hadoop can be successfully utilized for high-volume image files. Hence, eight different practical image processing algorithms are used for the utilization of Hadoop for image processing. Here Hadoop uses the extended file approach in to regard the whole TIFF image file as a unit by expanding the file format. Then, it is applied to other image formats such as the JPEG, BMP, and GIF formats. Experiments conclude that the method is scalable as well as efficient in processing large number of images used mostly for remote sensing applications, also the difference between the single PC runtime and the Hadoop runtime can be noticed.

[5] Recognition of Satellite Images of Large Scale Data Based on Map- Reduce Framework Vidya Jadhav, Harshada Nazirkar, Sneha Idekar, Prof. P.A. Bandgar, IJERA, ISSN: 2248–9622, Volume-4, Issue-3, March 2014.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

This paper suggests that for handling Images on large datasets, the utilization of parallel and distributed processing capability of Hadoop M-R is required. The current system of land-cover recognition gives a flexible solution for analysing satellite imagery automatically, even when GIS data is unavailable, or surface change may occur due to catastrophic events such as earthquake, flooding, hurricane, and snow storm, etc.

[6] Localized Image Retrieval Based on Interest Points Meng Fanjie*, Guo Baolong, Wu Xianxiang, SciVerse ScienceDirect, ISSN: 1877 – 7058, Volume-3, Issue-3, Elsevier 2011

This paper gives an ideal method based on interest points for content-based image retrieval. From the scale and rotation normalized image, interest points are to be detected. After that according to the distribution of interest points, image is divided into a series of sector sub-regions with different area. With robustness to the image's rotation, scale and translation, local features of every sector sub-region are extracted to describe the image and make the similarity measure. Images are regarded as multi-instance (MI) bags in the feedback phase, and the MI learning algorithm is employed to compute the target image feature, which results in calculating the similarity.

[7] System Refinement for Content Based Satellite Image Retrieval NourElDin Laban , Motaz ElSabanx, Ayman Nasr and Hoda Onsix Department of Image Processing and its Applications, Data Reception and Analysis Division, National Authority for Remote Sensing and Space Sciences (NARSS)

In this paper an easy approach that refines image retrieval process using the unique properties to satellite images is introduced. For the content of interest instead of using the more conventional rectangular query by image approach, it uses a Query by polygon (QBP) paradigm. Features are extracted from the satellite images using multiple tiling sizes. Accordingly the system uses these multilevel features within a multilevel retrieval system that refines the retrieval process. The multilevel refinement approach has been experimentally validated against the conventional one yielding enhanced precision and recall rates.

III. PROJECT IDEA

Satellite images play a major role in today's world in real-time event detection. These events may vary from changing landforms, depleting glaciers to catastrophic events like earthquakes, tsunamis and sand storms. The drastic changes after such events need to be monitored and capturing satellite images for such event detection can be helpful. The idea behind this project is to detect the changing landforms across different vegetation's, store this data, classify it on the basis of certain specified parameters and retrieve the classified data using well defined mechanisms. Segmentation and event detection is highly scalable in satellite images. With the increasing need to have real-time, classified data for specific applications there is an increasing need to store this chunk of data in a distributed environment to have better access. The basic idea is to capture the satellite images and store them in a distributed environment. The environment to be chosen is Hadoop Distributed Environment. Hadoop has the capacity to run applications on systems with thousands of multiple nodes involving thousands of terabytes of data. Its distributed file system not only facilitates rapid data transfer rates among nodes but allows the system to continue operating uninterrupted in case of a node failure as well. Once the captured images are stored, a training classifier will classify the images according to the vegetation, terrain, soil type and many such domains which will be considered according to the requirement. After classifying the data, retrieval of data will purely depend on the classification done by the Training classifier and desired result by the user. For this purpose we will be considering the MapReduce function to deduce the images from distributed datasets stored in the Hadoop environment to the user desired or factor specific images required. Further to this, a Running classifier will also run in parallel which will help in functions like zoom in, zoom out and distance calculation. The Running classifier will help the Training classifier to generate accurate result, as well as allow the user to do basic operation on the satellite images.

IV. PROPOSED ALGORITHM

Various Steps to be followed in the algorithm is as per following:

Algorithm 1:

- 1) Start
- 2) Convert color image to grayscale
- 3) Apply Canny edge detection algorithm



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

- 4) Calculate histogram
- 5) Compare with edge histogram of the database image
- 6) Stop

Algorithm 2:

- 1) Start
- 2) Acquire the satellite image
- 3) Convert color image to grayscale
- 4) Convert grayscale to binary
- 5) Count number of pixels in the image vicinity
- 6) Multiply pixel count with one pixel value
- 7) Compare with database image
- 8) Stop

Algorithm 3:

the following shows the canny edge detection algorithm steps.

- 1) Smoothing: Image will be blurred to remove the noise.
- 2) Find gradients: The edges are marked where the gradients of the image have large magnitudes.
- 3) Non-maximum suppression: Local maxima to be marked
- 4) Double thresholding: thresholding will determine Potential edges
- 5) Edge tracking by hysteresis: Suppress all edges that are not connected to a very certain (strong) edge to determine final edges.

Algorithm 4:

- 1) The canny operator works in a multi-stage process. Then a simple 2-D first derivative operator is applied to the smoothed image to highlight regions of the image with high first spatial derivatives.
- 2) Gradient is the first \hat{a} order derivatives of image for each direction. Which is non maximal suppression. The gradient can be computed using central difference.

$$\delta X(x,y)=[(x+1,y)-(x-1,y)]/2$$

$$\delta Y(x,y)=[(x,y+1)-(x,y-1)]/2$$

- 3) Magnitude of horizontal and vertical gradient is used.

The magnitude can be computed by

$$\text{Magnitude} = (\delta X(x) * \delta X(y) + \delta Y(x) * \delta Y(y))$$

V. PSEUDO CODE

S = {s, I, O, fn, fr }

where,

- s → Start state.
- I → Inputs from the given datasets (Satellite images or saved datasets).
- O → Output functions for displaying output to end user (web application).
- fn → Set of functions used for Image processing.
- fr → Set of related functions.

fn = { Is , If }

where,

- Is → ImageSegmentation()
- If → ImageFeatureExtraction()

Is =

f0 → Location(x-1, y-1)

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

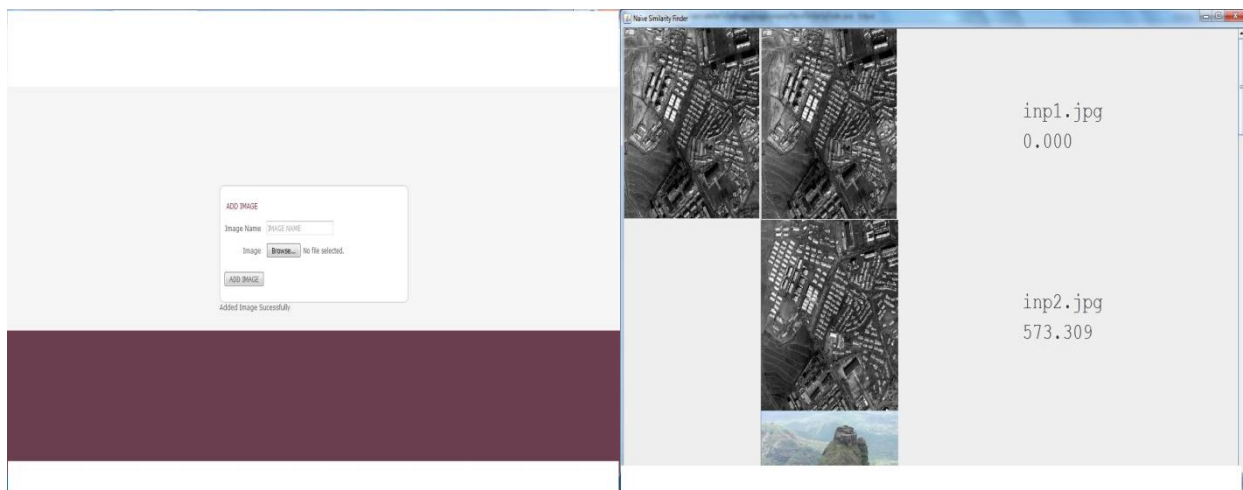
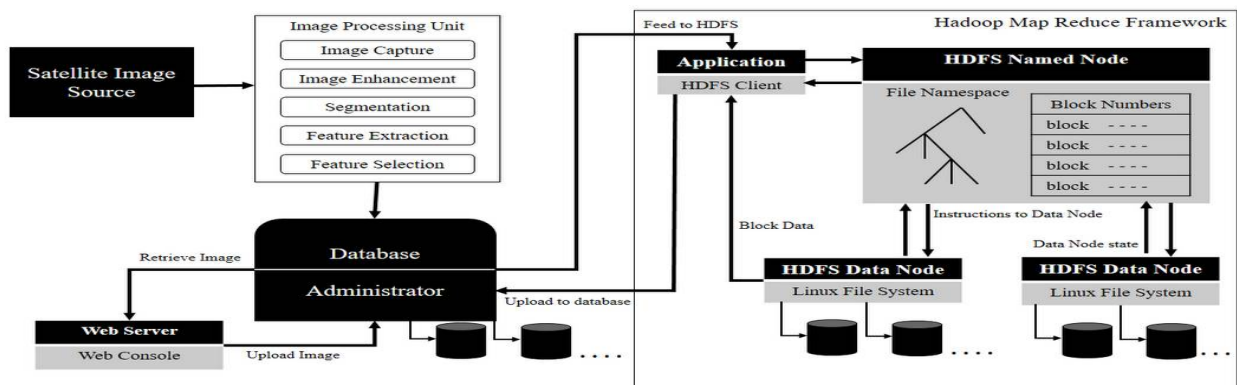
- f1 → Location(x, y-1)
- f2 → Location(x+1, y-1)
- f3 → Location(x-1, y)
- f4 → Location(x, y)
- f5 → Location(x+1, y)
- f6 → Location(x-1, y+1)
- f7 → Location(x, y+1)
- f8 → Location(x+1, y+1)

If = { fe0 , fe1 , fe2 , fe3 }

- fe0 → ShapeCheck()
- fe1 → SizeCheck()
- fe2 → ColourCheck()
- fe3 → TextureCheck()

VI. SIMULATION RESULTS

Overall architecture of proposed system is as shown in figure below. First part of the architecture will be satellite image extraction and later will be the Hadoop MapReduce.





International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

There are two main concepts 1. HDFS (Hadoop Distributed File System) 2. MapReduce.

We have divided the satellite images according to their contents in the respective nodes. The Fig.1 above depicts the insertion of images in database, and fig.2 gives the output. The final simulation results will also contain the tags of the respected images accordingly for user convenience.

VII. CONCLUSION AND FUTURE WORK

This system as of now can be more effective than the existing distributed system which is more complex and closed under the licence. The system can be extended in the military applications monitoring the Country borders or monitoring the catastrophic events in Weather Forecasting Applications too. Thus the project will implement an open source system to implement images retrieval using Hadoop MapReduce which gives the advantages over Existing System for distributed image processing i.e. Mat lab, which is closed under the licence and more complex system.

REFERENCES

- [1] Sarade Shrikant D., Ghule Nilkanth B., Disale Swapnil P., Sasane Sandip R., "LARGE SCALE SATELLITE IMAGE PROCESSING USING HADOOP DISTRIBUTED SYSTEM", IJARCT, ISSN : 2278 – 1323, Volume-3, Issue-3, March 2014.
- [2] Banaei, S.M. and Moghaddam, H.K. (2014) "Hadoop and Its Role in Modern Image Processing". Open Journal of Marine Science, 4, 239-245. <http://dx.doi.org/10.4236/ojms.2014.44022>
- [3] Penglin Zhang, Zhiyong Lv, and Wenzhong Shi., "Object-Based Spatial Feature for Classification of Very High Resolution Remote Sensing Images", IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, ISSN : 1545-598X, Volume-10, Issue-6, March 2013.
- [4] Mohamed H. Almeer, "Hadoop Map Reduce for remote Sensing Image Analysis", IJETAE, ISSN : 2250–259, Volume-2, Issue-4, April 2012.
- [5] Vidya Jadhav, Harshada Nazirkar, Sneha Idekar, Prof. P.A. Bandgar, "Recognition of Satellite Images of Large Scale Data Based on Map-Reduce Framework", IJERA, ISSN : 2248–9622, Volume-4, Issue-3, March 2014.
- [6] " Meng Fanjie*, Guo Baolong, Wu Xianxiang, "Localized Image Retrieval Based on Interest Points, SciVerse ScienceDirect, ISSN: 1877 – 7058, Volume-3, Issue-3, Elsevier 2011
- [7] NourElDin Laban , Motaz ElSabanx, Ayman Nasr and Hoda Onsix, "System Refinement for Content Based Satellite Image Retrieval", Department of Image Processing and its Applications, Data Reception and Analysis Division, National Authority for Remote Sensing and Space Sciences (NARSS)
- [8] Learn Map Reduce with Playing Cards: <https://www.youtube.com/watch?v=bcjSe0xCHbE>
- [9] Hadoop Fundamentals & Installation: www.tutorialspoint.com/hadoop/hadoop_tutorial.pdf
- [10] Intro to Hadoop and Map Reduce- How to Process Big Data (Udacity Course Material): <https://www.udacity.com/course/intro-to-hadoop-and-mapreduce--ud617>