View metadata, citation and similar papers at core.ac.uk brought to you by CORE
provided by MAT Journals

MAT JOURNALS
Journal of Computer Science Engineering and Software Testing
Volume 4 Issue 3

A Survey Paper on Satellite Image Using OpenCV Library over Hadoop Framework

Chaitanya Shrikant Kulkarni
Associate Professor,
Department of Computer Engineering,
VPKBIET, Baramati
Maharashtra, India
E-mail: *chaitanya.kulkarni@vpkbiet.org*

DOI: *http://doi.org/10.5281/zenodo.1476310*

Abstract

In this survey paper, we tend to study land classification from two-dimensional high resolution satellite pictures victimization Hadoop framework. Propelled picture process calculations that need higher system control with monstrous scale inputs is prepared with productivity exploitation the parallel and dispersed procedure of HadoopMapReduce Framework. HadoopMapReduce could be a climbable model that is fit for process petabytes information with enhanced adaptation to non-critical failure and information closeness. During this paper we tend to gift a MapReduce framework for acting parallel remote sensing satellite processing victimization Hadoop and storing the output in HBase. The speed and performance show that by utilizing Hadoop, we are able to distribute our employment across completely different clusters to require advantage of combined process power on goods hardware.

***Keywords**: Hadoop Map-Reduce,Hbase, Opencv*

INTRODUCTION

We ponder a framework dependent on Hadoop an open source framework that actualizes the MapReduce programming model and that can enhance the classification of extensive scale remote detecting picture and benefit the intensity of spatial huge information idea.

Hadoop is an open source distributed framework based on Google's MapReduce processing technique and distributed file system framework. Hadoop framework mainly consists of Hadoop Distributed File System (HDFS) and HadoopMapReduce.

Knowledge in Hadoop cluster is split into smaller items and distributed throughout the cluster. The principal target of HDFS is to store learning reliably even in nearness of disappointments together with Namenode disappointment, information hub disappointment and disappointment on account of system segments. MapReduce could be a programming model intended for process monster volume of information in parallel by separating work into set of individual assignments. The most task of MapReduce is to separate computer file set into freelance chunks that are processed in fully parallel manner. Then processed images are stored in HBase. HBase is a distributed database framework that is generally utilized when there is a need for random, real-time read/write access to large table.

The study done by several researchers counsel that classification is accomplished by exploitation one variable i.e., spectral characteristic (colour) or black and white tone. However, few researchers have used appurtenant information (such as DEM) together with remote sensing information to enhance classification accuracy. Still alternative variable like patterns, association, scale, texture etc area unit needed to lean importance in classification to extend the accuracy. For alittle region, the classification are terribly simple because the range of classes area unit less and field verification might be done, however once a district of study multiplied the classification accuracy additionally decreases because of increment in classes gift & field verification is additionally not possible for every purpose.

The purpose of this survey is to giant datasets of a high resolution satellite image classified and performance achieved victimization hadoop framework. The populated areaof Pune has been hand-picked for this purpose because it includes comparatively advanced land cowl and land use patterns that are appropriate for classification algorithmic program comparison. Additionally to applying the algorithms on a pixel-by-pixel basis, we tend to additionally tested the algorithms on a per-segment basis to check the impact of together with the object-based image analysis as a pre-processing step within the image classification method.

In this work, we tend to analyze a framework to with efficiency method extremely voluminous satellite knowledge victimization Hadoop. We tend to propose an optimized HDFS-Hadoop framework methodology with simple application programming interface that may with efficiency method high resolution satellite image knowledge in gait.

HDFS OPTIMIZATION

In the following, we briefly discuss the terminologies of Hadoop, HDFS and Map-Reduce.

HADOOP

HDFS and MapReduce: Hadoop Distributed classification system (HDFS) will store massive files (typically within the vary of gigabytes to terabytes) across multiple machines. It achieves dependableness by replicating the information across multiple hosts. Master-Slave design is employed in HDFS. Master is thought as Namenode and slave is named Datanode. Namenode stores solely data regarding the classification system. Application information is hold on on multiple Datanodes within the type of 64MB blocks. The Map-Reduce engine may be a parallel programming technique that runs on high of the HDFS. Map-Reduce engine consists of 1 JobTracker to that shopper applications submit MapReduce jobs. The JobTracker then assigns the task to offered TaskTracker nodes within the cluster, endeavour to stay the work as on the point of the information as attainable (Data locality).

HDFS Optimization

When huge numbers of small files (order of KiloBytes) need to be stored in the HDFS, memory usage in Namenode (Master Node) increases. There is a fixed overhead involved in storing every file, directory or block in HDFS. This overhead contains meta-information related to the stored data like file and data locations. This overhead is represented as an object in the Namenode’s memory, each of which requires some amount of storage, say 160 bytes. So, if there are 10 million files, each using a HDFS block, would consume around 4 GigaBytes(GB) of memory (considering two backup copies of the data stored in different nodes). Handling such large volume of data renders the master node unresponsive for servicing file operation requests from client nodes. Certainly, handling a billion files is not at all feasible. Furthermore, reading huge number of small files causes lot of file seeks and lots of hopping from one Datanode to the other to retrieve each small file. All of these result in an inefficient data access pattern that leads to poor performance. Handling small files also affects MapReduce performance.

Usually, a Map task processes a block of input at a time. If the file is very small, then each map task processes very little input. To handle many such small files, there need to be many more Map tasks, each of which contributes to extra bookkeeping overhead. For example, consider splitting of a 1GB file into sixteen 64MB files vs 10,486 100 KB files. Each of the 10,486 files needs one Map task each. This results in the overall job time being tens or hundreds of times slower than the equivalent processing of one larger file (64 MB). Dong et al. in [8] proposed two techniques for solving this problem. First technique is file merging and pre-fetching scheme for structurally-related small files. Second technique is file grouping and pre-fetching scheme for logically related small files. Their approaches are based on categorization of files based on their logical or structural properties. In this paper, we propose similar grouping strategies that can apply for processing high resolution image datasets, thereby optimizing storage space on HDFS. These HDFS optimization techniques thus help to reduce network overheads, and can drastically reduce overall processing time. This is in addition to a Map-Reduce implementation that resolves out of memory issues while processing highresolution images.

Many Image processing tools are available where we can do Image Processing. But the main drawback is that these software and tools are optimized for single machine and are sequential in nature. If we do the processing of Remote sensing data in the sequential manner (image after image) it will take long time to process huge number of high resolution remote sensing images. Even techniques like thread parallelism or batch processing have their own hardware constraints and reliability issues. Hence, we require a multimode framework that can employ parallelism in the most efficient manner and ensure guaranteed processing of each datum. This is provided by Hadoop. Also, the principle of Hadoop Implementation is in moving the computation to the processing node rather than moving data, thereby ensuring data-locality. In the case of processing high resolution remote sensing satellite images, the size of input data is much larger than the computation involved. Based on these facts, one can conclude that Hadoop framework is optimally suited for this job.

IMAGE PROCESSING

OpenCV

OpenCv is relate open supply PC vision library containing more than five hundred upgraded calculations for picture and video investigation, and also manufactory item examination, therapeutic imaging, security, PC program, camera movement, stereo vision and computerized reasoning. Since its presentation in 1999, it's been generally received in light of the fact that the essential advancement instrument by the network of analysts and engineers in PC vision. The library is composed in streamlined C and exploits multi-center processors. Rendition 1.0 was propelled in a couple of006 and second major unharness happened in 2009 with dispatch of OpenCv 2 that anticipated fundamental changes, especially the new C++ interface. This new interface looks for scale back the measure of lines of code important to code up vision reasonableness moreover as lessen regular programming mistakes like memory spills (through programmed data assignment and deallocation) which will emerge once exploitation OpenCV in C. As of late, the chief of late improvements and calculations in OpenCV are created inside the C++ interface. There's dynamic advancement on interfaces for:

1. Python.
2. Ruby.
3. Mat lab.
4. And other.

At the time of writing, the latest release is 3.2 (December 2016). A graphical interface can be created by Qt cross platform application and UI framework with APIs for C++ programming and Qt Quick for rapid UI creation.

Working with Multi-Processor Station

With the end goal to enhance proficiency of multi-center PC utilize, it is important to compose advanced codes. Utilizing more strings or laborer. The laborers in Matlab are given by capacity matlabpool. This capacity opens and closes pool of Mat lab session for parallel calculation. Matlabpool empowers the parallel dialect includes by beginning a parallel employment that interfaces this Mat lab customer with number of specialists. Greatest number of laborers is determinated by number of processors. Multi-threading in OpenCv can be accomplished in various ways. The Threading Building Blocks (TBB), offered by Intel, is one of the ways. This arrangement offers a rich and finish way to deal with communicating parallelism in a C++ program . Other way is use of OpenMP (Open Multi-Processing). In this work is used another way, following POSIX (Portable Operating System Interface) threads standard. It's enforced with a pthread.h header and a thread library. The first argument may be a pointer to prhread_t. Second sets thread attributes, in most case, this argument is ready to NULL. Next argument may be a thread operate that's capital punishment with thread begin. Last argument contains information, that are passed to the thread operate. Once a thread terminates, the pthread_exit operate is named. In main method, pthread_join may be operate, that processes use to gather kid processes.

CLASSIFICATION FUNTIONS

Smoothing

Smoothing could be a straightforward and often used image process operation. There square measure several reasons for smoothing, however it's typically done to cut back noise, camera artifacts or facilitate finding edges. OpenCV offers 5 completely different smoothing operations, that square measure supported through one operate:
*voidcvSmooth(*src, *dst, smoothtype, p1, p2, p3, p4).*

Changing Colorspaces

There are over a hundred and fifty color-space conversion strategies on the market in OpenCV., to convert pictures from one color-space to a different, like BGR to grey, BGR to HSV etc.In addition thereto, we have a tendency to produce associate application that extracts a coloured object during a video OpenCv provides operate: cv2.cvtColor(), cv2.inRange() etc. For color conversion, OpenCv provides function: cv2.cvtColor(input, image, flag)where flag determines the type of conversion. For BGR to Gray conversion we use the flags cv2.COLOR_BGR2HSVGRAY. Similarly for BGR to HSV, we use the flag cv2.COLOR_BGR2HSV.

Image Thresholding

The simplest thresholding strategies replace every picture element in a picture with a black picture element if the image intensity Iij is a smaller amount than some mounted constant T (that is, Iij< T), or a white picture element if the image intensity is bigger than that constant. cv2.threshold, cv2.adaptiveThreshold etc. There square measure several thresholding sorts we are able to replace the picture element intensity.

Simple Thresholding

Basic thresholding is straight forward. On the off chance that pixel esteem is more prominent than a limit esteem, it is allocated one esteem (might be white), else it is appointed another esteem (might be dark). The capacity utilized is cv2.threshold. First contention is the source picture, which ought to be a grayscale picture. Second contention is the limit esteem which is utilized to order the pixel esteems. Third contention is the maxVal which speaks to the incentive to be given if pixel esteem is more than (in some cases not as much as) the edge esteem. OpenCV gives distinctive styles of thresholding and it is chosen by the fourth parameter of the capacity. Distinctive composes are:

1. cv2.THRESH_BINARY
2. cv2.THRESH_BINARY_INV
3. cv2.THRESH_TRUNC
4. cv2.THRESH_TOZERO
5. cv2.THRESH_TOZERO_INV

Two outputs are obtained. First one is a retval which will be explained later. Second output is our thresholded image**.**

Adaptive Thresholding

Versatile thresholding is utilized a worldwide incentive as edge esteem. However, it may not be great in every one of the conditions where picture has distinctive lighting conditions in various regions. All things considered, we go for versatile thresholding. In this, the calculation ascertains the edge for a little areas of the picture. So we get distinctive limits for various districts of a similar picture and it gives us better outcomes for pictures with differing illumination.It has three 'extraordinary' input params and just a single yield contention.

1. cv2.ADAPTIVE_THRESH_MEAN_C :limit esteem is the mean of neighborhood region.
2. cv2.ADAPTIVE_THRESH_GAUSSIAN_C:limit esteem is the weighted whole of neighborhood esteems where weights are a gaussian window. Square Size - It chooses the span of neighborhood territory. C - It is only a steady which is subtracted from the mean or weighted mean ascertained.

Edge Detection

Edge location is an essential pre-preparing venture in picture examination. By and large, the most widely recognized administrator used to speak to edges is the Sobel subordinate administrator. OpenCV gives work

*cvSobel(*src, *dst, xorder, yorder, aparature_size).* Contentions src and dst are information and yield picture, xorder and yorder are the requests of derivate. Normal qualities are 0,1 or at generally 2. The aperture_size parameter ought to be odd and decide the width of the square channel. Sobel subsidiaries have the decent property that they can be characterized for portions of any size, and those parts can be built rapidly and iteratively. The bigger portions give a superior estimate to the subsidiary on the grounds that the littler parts are extremely touchy to clamor. Keen edge identifier is an edge area executive that uses a multi-sort out count to recognize a broad assortment of edges in pictures. The principal subordinates are figured in x and y and afterward joined into four directional subsidiaries. The focuses where these directional subordinates are neighborhood greatest are then contender for amassing into edges.OpenCv gives work: void cvCanny(*img, *edges, low_thresh, high_thresh, opertureSize). Capacity expects an info picture, which must be grayscale, and a yield picture, which must be additionally grayscale. The following two contentions are the low and high edges, and the last contention is another gap. Not surprisingly, this is the gap utilized by the Sobel subsidiary administrators that are called within the usage of cvCanny().

STUDY AREA

The area of study, JSPM NTC,Pune which is Engineering college in Pune, Maharashtra lies in east of Mumbai and north of Satara and extends between Lat.

18° 26' 30" N and Long. 73 ° 45′E to 73° 49' 55" E (Fig.1). It covers an area of 29.56 (acres). The topography of the study area is not very complex.

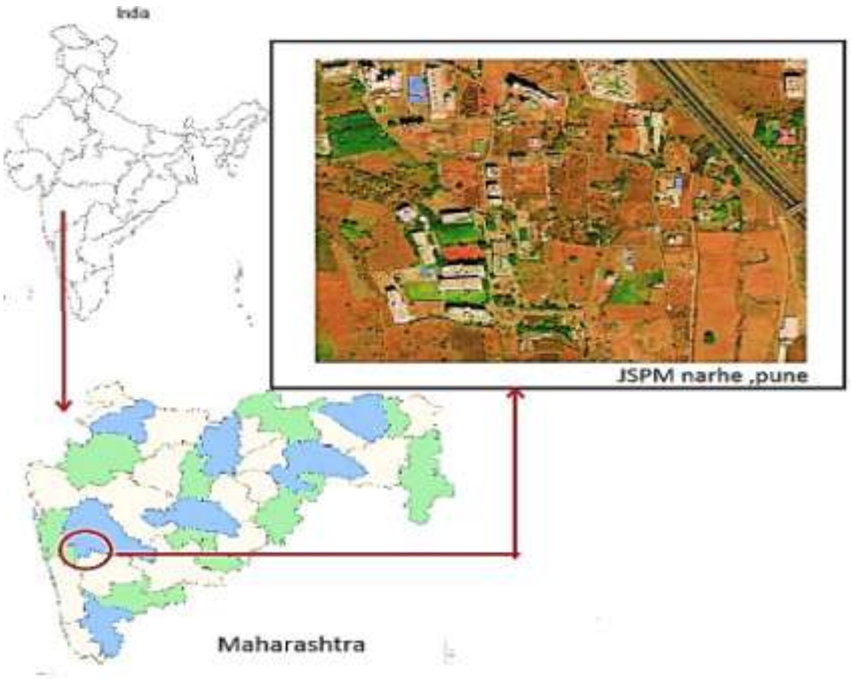

Fig. 1: *Location of the study area.*

SATELLITE IMAGERY

Satellite symbolism comprises of pictures of Earth or different planets gathered by satellites. Imaging satellites are worked by governments and organizations around the globe. Satellite imaging organizations offer pictures under permit. Pictures are authorized to governments and organizations, for example, Apple Maps and Google Maps. Followings are open website application for downloading satellite images:

1. https://earthexplorer.usgs.gov/
2. https://www.daftlogic.com/projects-google-maps-area-calculator-tool.htm
3. http://bhuvan.nrsc.gov.in/data/download/index.php

SASPlanet is a program intended for review and downloading high-goals satellite symbolism and customary maps presented by such administrations as Google Maps, DigitalGlobe, Kosmosnimki, Yandex.Maps, Yahoo! Maps, VirtualEarth, Gurtam, OpenStreetMap, eAtlas, Genshtabmaps, iPhone maps, Navitel maps, Bings Maps (Bird's Eye) and so forth., however as opposed to every one of these administrations all downloaded pictures will stay on your PC and you will have the capacity to see them, even without associating with the web.

CONCLUSION

This work demonstrates the efficiency of by using Hadoopcluster framework for processing high resolution satelliteimage data with optimized HDFS storage. This study is aneffort to address core issues in remote sensing such as need for large computational demand. Although we considered remote sensing satellite data for our implementation, the image processingcan be used for classifying land cover such as green land, bare land, water surface, constructed area. The results show that proposed Opencv library provides significant classification of a high resolution image data.

REFERENCES

1. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.
2. http://static.googleusercontent.com/media/research.google.com/es/us/archive/mapreduce-osdi04.pdf.
3. http://wiki.apache.org/hadoop/Hbase
4. Mohamed H. Almeer, "Cloud Hadoop Map Reduce For Remote SensingImage Analysis" Journal of Emerging Trends in Computing andInformation Sciences,Vol. 3, No. 4, April 2012
5. H Kocakulak and T TTamizel, "A Hadoop solution for ballistic imageanalysis and recognition" International Conference on High PerformanceComputing and Simulation (HPCS), Istanbul, pp. 836-842, 2011.
6. B. Li H Zhao, Z H Lu, "Parallel ISODATA clustering of remote sensingimages based on MapReduce", International Conference Cyber enableddistributed computing and Knowledge Discovery (CyberC), Huangshan,pp. 380-383, 2010.
7. Z. Lv, Y. Hu, H. Zhong, J. Wu, B Li and Z Zhao, 2010 "Parallel Kmeans clustering of remote sensing images based on MapReduce",International Conference on Web Information Systems and Mining, pp.162-170. 2010.
8. Dong, B. "An Optimized Approach for Storing and Accessing SmallFiles on Cloud Storage", Journal of Network and ComputerApplications, 1847-1862, 2012.
9. Anderson J.R., Hardy E.E., Roach J.T. and Wirmer R.E., "A land useand land cover classification system for use with remote sensing data,"U.S.G.S. Professional paper No. 964 USGS Reston Virginia, pp. 1-28,1976.
10. Chopra Narayan &Prudhvi Raju K.N., "Visual Interpretation withRemote Sensing Data of Coarse Resolution: Some Points to Ponder,"Indian Journal of Landscape Systems and Ecological Studies, vol. 33(1), pp. 35-38, 2009.
11. *OpenCV Reference Manual*, v 2.1, march 18, 2010
12. *MathWorks documentation*,http://www.mathworks.com/help/toolbox/distcomp/matlabpool.html , available online on 20.3.2012,
13. OpenCv on-line documentation,http://opencv.willowgarage.com/wiki/TBB , available online on 20.3.2012
14. SlavomirMatuska, Robert Hudec and MiroslavBenco-"The Comparison of CPU Time Consumption for Image Processing Algorithm in Matlab and OpenCV"- 978-1-4673-1179-3/12/$31.00 IEEE 2012
15. Xinghui Dong and Mike J. Chantler-"Perceptually Motivated Image FeaturesUsing Contours" IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 25, NO. 11, NOVEMBER 2016

Cite this article as: Chaitanya Shrikant Kulkarni. (2018). A Survey Paper on Satellite Image Using OpenCV Library over Hadoop Framework. Journal of Computer Science Engineering and Software Testing, 4(3), 12–18. http://doi.org/10.5281/zenodo.1476310