Al-Salam Journal for Engineering and Technology
Journal Homepage: http://journal.alsalam.edu.iq/index.php/ajest
ISSN: 2790-4822

Image mining technique using Hadoop map reduce over distributed multi-node computers connections

Zahraa Azhar Muhammad Shamki[1], Furkan Rabee[2*]

[1]University of kufa Al-Najaf Al-Ashraf, Iraq

[2]University of kufa Al-Najaf Al-Ashraf, Iraq

DOI: https://doi.org/10.55145/ajest.2022.01.02.004
Received May 2022; Accepted June 2022; Available online July 2022

ABSTRACT: The amount of today's image collections has exploded to petabytes of data. A fair amount of time cannot be allotted for the analysis of such massive datasets using a personal computer. As a result, distributed computing is required for current image collection mining. This work used multi-nodes-computers for image mining in order to improve reliability, fall tolerance, and time efficiency. Because of this, the data was divided up across the nodes in the Hadoop multi-node cluster, and the results were then compiled to create an image clustering algorithm. One master node and two slave nodes were used to test our technique on a huge dataset. We found that Using multi-node Hadoop (48%) faster than traditional image mining implementation, and (24%) faster than single-node Hadoop implementation.

Keywords: image mining, K-Mean algorithm, Multi node Hadoop, feature extraction.

1. INTRODUCTION

As computer science advances rapidly and the Internet's popularity keeps rising, the Internet has become a center for exchanging any and all data. Everybody is posting and sharing photos on Facebook and sending E-mails [1]. New and useful information may be gained from these images using image mining [2]. Image mining is the process of extracting information from images that is not immediately apparent, such as patterns or correlations. Implicit knowledge, image data exchanges, and other kinds of connections can be included [3]. Images may be mined in two ways using these procedures. Extracting data from databases or collections of photos, and mining a mix of alphanumeric data and collections of photographs [4]. Patricia G. Foschi discusses that Feature selection and extraction are the first steps in image mining's pre-processing, Obviously this is a critical step in the entire scenario of Image Mining [4] Interacting with images and extracting relevant information from images with descriptions based on the attributes of the images itself is referred to as "feature extraction." [5]. In order to do research, these features must be clustered together. Using image characteristics such as shape, texture, and color, the image mining process of clustering organizes images into groups[6] a lot of time is lost because these images are being analyzed and mined with traditional processing methods[1] The Hadoop framework is used to handle these images on a large scale, making it easier and faster. Using a multi-node Hadoop cluster, map-reduce processing increases efficiency, resulting in time savings.

1.1. Apache Hadoop

Open-source software for reliable, scalable, distributed computing is made by the Apache Hadoop project. A Hadoop system typically has one master node in addition to a number of worker nodes. The master node, which is known as NameNode, is responsible for managing the request jobs that are sent in by users, breaking down each job into many tasks, and then delegating those tasks to different workers. The workers that are known as DataNodes are the ones in charge of storing data and carrying out the activities that have been planned by NameNode [7].The Apache Hadoop software lets clusters of computers work together to process large sets of data in a simple way. It lets applications work with petabytes of data and thousands of computers that can do their own work. Hadoop came from Google's MapReduce and Google File System (GFS) [8]. The processing component of Hadoop is referred to as MapReduce, while the storage component is referred to as the Hadoop Distributed File System (HDFS)[9]. Hadoop is built on a master/slave design. There is a master and a lot of slaves in this architecture. Master runs the NameNode,

*Corresponding author: zahraa.alramahi@uokufa.edu.iq

JobTracker/MapReduce, and TaskTracker/MapReduce, while slave runs the DataNodes/HDFS and TaskTracker/MapReduce for data storage. When a client asks anything from NameNode, it is handled by NameNode. If you're working with data stored in HDFS, you'll find that the NameNode is responsible for keeping track of all storage-related metadata [10]. A "Hadoop cluster" is a "computational cluster" that is made up more than one machine linked together using the Hadoop framework. It is mostly used to store and examine massive data, like images, text files, videos, and other[1].So Hadoop architecture may be configured as either a single-node cluster or a multiple-node cluster

1.1.1. Single node Hadoop cluster

All daemons with a single-node cluster can be installed and configuring on a one virtual machine. This strategy is typically employed during the investigation and testing phase or in contexts with little data, but Hadoop might not be the best way to store data in this case. If you don't have enough data, most of Hadoop's main benefits won't be clear right away [10].

1.1.2. Multi node Hadoop cluster

Using several simulated machine Setup of Hadoop in a multi-node cluster requires virtualization. To put it another way, each data node is running on its own virtual machine. For BigData analysis, enterprises employ this type of infrastructure. For real-world use, you need to distribute petabytes of data among hundreds of machines in order for it to be processed instantly [10].

1.2. MapReduce Framework

MapReduce is a Java-based software framework that lets programmers run the same computation on multiple machines at the same time to process data faster, reliable, fault-tolerant manner and more efficiently. Apache Hadoop is widely used in science because it works well with the MapReduce programming language and is free to use [11, 12]. With the help of the Google File System (GFS), the information is broken up into smaller pieces and sent to many computers. Using MapReduce, which is a parallel programming API, calculations are distributed to the data (thus the name "Map") and combined at the end (called "Reduce") [13]. These two functions are done at the same time; data is saved as pairs of "key, value." Map function starts by reading a value from the input file. This value is then applied to the function, yielding intermediate output values. These intermediate results are likewise stored in the cluster nodes as key-value pairs. Any key's records could go through many nodes. Sorting is done on the output of the map function and then sent to the reduce function [14]. The MapReduce algorithm works well for mining petabyte-sized datasets that can't be stored physically [14] .

The rest of this study is broken down into two sections: the second portion discusses previous research, and the third section discusses proposed research. Section 4 discusses the experiment's findings and implications.

2. Related Work

Numerous researches have looked at image mining. Earlier image mining approaches are summarized in this portion of the text. Image mining is a fundamental method for gleaning information from visual data. An image processing application of data mining is all that's required [15].

Reference [16] Color is a differentiating feature which uses Block Truncation Coding (BTC) and Color Moment (CM) to get features of image dataset. The image collection is then divided into various groups using the K-Means clustering algorithm.

Handwritten signature feature extraction and handwritten signature recognition were recently introduced by (Biswas1et.al. 2010) that developed a way to extract characteristics from Handwritten Signature Images. That calculated characteristic is utilized for verification. Here we employed a clustering approach for verification. As a result of the research presented here, an image clustering method based on a k-nearest neighbor’s technique has been developed that is capable of handling clusters of various sizes and shapes. This approach is obviously capable of distinguishing forgeries from actual images, as demonstrated by the results of experiments [17]. [18] Proposed that Clustering is one of the most widely used image mining methods, it may be used in a variety of fields, including image segmentation and bioinformatics. As a result of its ease of implementation, simplicity, efficiency, and empirical success, K-means has become the most common and simple clustering method. However, real-world applications generate enormous amounts of data, and thus, how to properly manage this data in an essential mining operation has been a substantial and demanding problem. Additionally, as a message-passing programming model, MPI (Message Passing Interface) offers great performance, scalability, and portability. MKmeans is an MPI-based parallel K-means clustering technique that was inspired by this. The approach provides successful use of the clustering algorithm in a parallel context. Experimentation reveals that MKmeans is very robust and portable, and that it operates with minimal time overhead on huge data sets.

In the field of RS (remote sensing) images, deforestation, climate change, ecosystem and land surface temperature are some of the main research areas, where features need to be classified or clustered to provide research basis. Clustering using the K-Means algorithm is a fundamental technique used in the analysis of real-time RS images. Processing a huge number of RS images becomes impossible for PCs because of their limited hardware resources and their tolerance for long processing times. Parallel and distributed computing approaches are unquestionably the best options. Different with traditional ways, in [19] this approach was parallelized using Hadoop, an open source system that uses the MapReduce programming model to store and analyze big datasets ranging in size from gigabytes to petabytes, instead of the more traditional methods. In their research, they have shown that the outcomes are acceptable and may inspire new techniques to solving similar issues in remote sensing applications.

In the Big Data era, there are a lot of images, which makes it hard to find satellite images. High processing speed is now essential for some unique applications, like responding quickly to disaster warnings. Using K-Means clustering on the Hadoop system, we demonstrate an efficient method for detecting satellite images in [20]. They design the effective K-Means algorithm based on MapReduce programming model and Hadoop distributed file system. Two main operations in MapReduce: Map and Reduce, are realized to give an efficient implementation. The results show a fast detection speed and good scale up while keeping accuracy both in training and testing. For analyzing a vast number of photographs of fingerprints that could not otherwise be processed owing to a lack of physical memory, in [14] they turned to the MapReduce programming technique. Preprocessing and extraction of biometric features from images are done in an image data store before they are stored in a database. Multi-fingerprint images are preprocessed and extracted simultaneously by the algorithm in order to extract features (ridges and bifurcation). The results reveal a significant reduction in the amount of time it takes to generate a feature vector for each image processed.

Our work is combined between feature extraction method explained above by some researcher and image clustering this make image mining system in addition the proposed work applied on multi node Hadoop framework to obtain fast results

3. Proposed Work

In this paper, we talk about how image mining can be done in parallel on Hadoop using above mapreduce architecture. We use multi node Hadoop cluster as framework to apply the algorithm on it. Fig.1 shows how we use the Hadoop multi node cluster to mine images. The images in the dataset are first turned into a file that is stored on HDFS. In the second step, extract color and texture feature from images by using the Map and Reduce function on a file that is stored on HDFS; The Hadoop MapReduce job divides the input file into chunks each of size 64MB. The Map tasks on different nodes (PC) then work in parallel to gather all of the pixels for each image from these separate chunks. After sorting, the outputs are sent to the reduce tasks, which compute the color and texture features as shown in equations below and write the feature vector to an HDFS file.

Image mining is the process of finding features, patterns, and knowledge in large groups of images that are related to a certain domain. On the other hand, color, texture, and shape are available [4].Feature extraction is the main core in diagnosis, classification, clustering, recognition, and detection [21].

3.1. feature extraction

3.1.1. color feature

Color feature is one of the most important features of an image. It is defined to a particular color space or model [22]. This paper used color moment (CM) as method to extract color feature which it is one of the most straightforward yet powerful features. Mean, standard deviation, and skewness are the common moments, and the associated computation is as follows [21, 22]

1- Mean

It is the average color value of the image and may be determined using the equation (1) below[21]

$$M_j = \sum_{i=1}^{m} \frac{1}{m} Q_{ji} \tag{1}$$

2- Standard deviation

It is a measurement of the variance of a set of numbers. SD may be calculated using the square root of distribution variance; equation (2) describes its format [21]

$$\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Q_{ji} - M_j)^2} \tag{2}$$

3- Skewness

Equation (3) describes deviation as a measurement of asymmetry of the distribution [2] as shown in equation (3) [21]

$$S_j = \sqrt[3]{\frac{1}{m}\sum_{i=1}^{m}(Q_{ji} - M_j)^3} \tag{3}$$

where Qji is the value of pixel at location ji and m is the number of pixels in the image. Mean (E), Standard Deviation (SD), and Skewness (S) are used to make the feature vector of color representation: F_{Color} = (M, SD, S).

3.1.2. Texture feature

Texture feature extraction is a method for calculating and defining the features and attributes of an image that quantitatively represents the texture image's features **[23].** Gabor filter is used in this approach to extract texture feature, it is one of the most well-known feature descriptor that Gabor invented in 1946**[23]**. The Gabor filter consists of a Gaussian kernel function that is modulated by a complex sinusoidal plane wave[24], as illustrated in (4). It works well in both the frequency and spatial domains [25] and many different sizes **[23].**

$$f_{mn}(k,l) = gu_{mn}(k,l)s_{mn}(k,l) \tag{4}$$

Gaussian component gu(k,l), and the sinusoidal component s(k,l) is given in (5) and (6) respectively:

$$gu_{mn}(k,l) = \frac{1}{2\pi\sigma_m^2}\exp\left(\frac{(k^2 + l^2)}{2\sigma_m^2}\right) \tag{5}$$

$$S(k,l) = \cos\left(2\pi\left(U_m k\cos\theta_n + U_m l\sin\theta_n\right)\right) \tag{6}$$

k and l are the spatial variables, whereas m and n are the scale and orientation indices, respectively. This experiment used four scales (1,2,3,4) and eight orientations θ. Consequently, the orientations range from angle 0° to 157.5° with an “orientation bandwidth of B = 22.5°”. U1 in Equation (7) is utilized to identify the filter bank with the greatest spatial frequency. The parameter σ_m can be written in (8):

$$U_m = 0.5/\left(1 + \tan\left(B_\theta/2\right)\right) \tag{7}$$

$$\sigma_m = sqrt\left(2\ln 2\right)/\left(2\pi U_m\tan\left(B_\theta/2\right)\right) \tag{8}$$

Apply Gabor filter f_{mn} (k, l) to an image I(k, l) of size R x Q resulted discrete Gabor wavelet transform of an image (9):

$$G_{mn}(k,l) = I(k,l) * f_{mn}(k,l) \tag{9}$$

After applying Gabor filters to an image array at various scales and orientations, the energy content is computed in (10)

$$E(m,n) = \sum\sum|G_{mn}(k,l)| \tag{10}$$

If you want to cluster similar textures of images or regions, you may use the “mean μ_{mn}“ shown in (11) and “standard deviation σ_{mn}“ given in (12) to represent the region's homogeneous texture feature:

$$\mu_{mn} = E(m,n)/(R\times Q) \tag{11}$$

$$\sigma_{mn} = sqrt\left(\sum\sum|G_{mn}(k,l)| - \mu_{mn}\right)/\left(R\times Q\right) \tag{12}$$

$F_{Texture} = (\mu_{00}\ , \sigma_{00},\ \mu_{01}, \sigma_{01}, \ldots\ldots\ldots\ldots\ \mu_{mn}, \sigma_{mn})$

3.2. Parallel K-Means Algorithm with MapReduce on Hadoop

In the Hadoop multi node MapReduce framework, the extracted features F_{color}, $F_{texture}$ are clustered using a parallel K-Means algorithm with map reduce programming language. This method worked well, and it took less time than the normal sequential method and the Hadoop single-node method. The MapReduce jobs will be assigned to each iteration of the parallelized K-Means method: For a Map job, compute distance between each feature vector and each cluster center, then update cluster center in reduce task [13]. Records of data of dataset features are saved in rows in order to initiate the Map job. As a result, every Map job has a record of completion, and MapReduce on the Hadoop system automatically completes the operation [13].

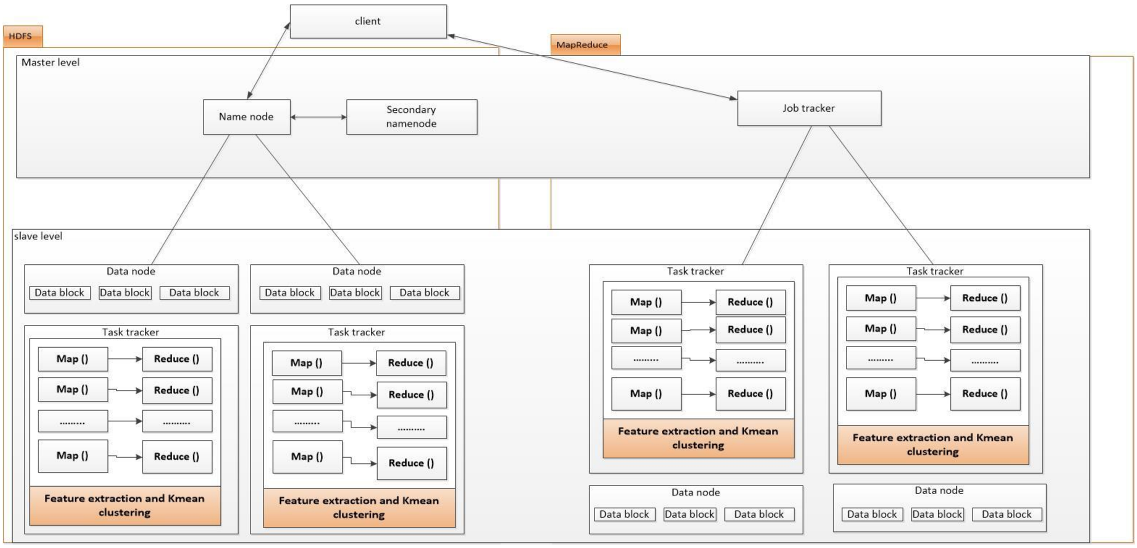

FIGURE 1. Image mining in multi node Hadoop cluster

```
Algorithm 3.1
Input: An RGB image dataset
Output: clusters of images
Method:
Step 1:begin
Step 2:put the file that contain pixels of all color image on HDFS
Step 3: divides the file in to chunk of size 64 MB
Step 4: distribute these chunks among Datanodes
Step 5: in each Datanode the work in map reduce is:
Step 6: for each image do:
        Step 6.1: begin
        Step 6.2: for each color component (R, G, B) in image do:
               Step 6.2.1: begin
               Step 6.2.2: calculate  (mean, standard deviation, skewness)
               Step 6.2.3: apply Gabor filter on color space
               Step 6.2.4: construct the feature vector of that color
               Step 6.2.5: compute the average feature vector of three colors
               Step 6.2.6:End
      Step 6.3: Write the feature vector $F_{color}$,$F_{Texture}$ on output file on HDFS
      Step 6.4: End
Step 7: apply K-mean clustering on this HDFS feature file in map reduce too
Step 8:Write the output file on HDFS
Step 9: end
```

3. Experimental Work and Results

We have used K-Means clustering on the "multi-node Hadoop cluster" to set up a method for mining images.

- The experiments' program is written in the Eclipse Java programming language and executed it in sequential way and single node Hadoop and multi node Hadoop.
- "Apache Hadoop 3.3" serves as a cluster-distributed computing framework in single node Hadoop.
- For the practical implementation of multi node, we used one master and two slaves in two Personal Computer with “Ubuntu Desktop 20.04.1 LTS” operating system
- “Apache Hadoop 2.7.1” installed on these nodes
- One of the PC acts as master another one acts as two slave nodes, “VMware Workstation 16 Pro” is used to create virtual machines.
- The component packages of the programming framework that was used for the experiments provide access to MapReduce and HDFS.
- Dataset used contains images whose pixels migrate to a file stored on HDFS. The database has three classes of 256×256 pixel images: rose images, berry images, and dog images.
- Large size of images file is 968 MB of 700 images while medium size of images file is 408 MB of 300 images and small size of images file is 132 MB of 100 images.
- In the first part of the valuation process, “the color moment and Gabor measurements” for each of three color components (Red, Green, and Blue) are worked out.
- The feature vector of each color layer of 67 elements, in which first three numbers is the mean, standard deviation, and skewness for a color feature, as well as thirty two means (μ_{mn}) and thirty two standard deviations (σ_{mn}) elements for Texture feature.
- The average feature vector for the whole image was worked out based on the feature vectors of each color layer.
- The second part of the plan is to use the “Parallel K-Means Algorithm” Based on the MapReduce Model on Hadoop System on the file that comes out of the map-reduce feature extraction program on HDFS.
- Experiment and results allow us to make comparison of execution time between multi-node Hadoop system, single-node Hadoop system and the conventional codes.
- Table 1 demonstrates that the feature extraction and parallel K-Means approach in multi-node Hadoop are quicker than the conventional clustering algorithm and single-node Hadoop. The findings are pretty satisfactory.

Table 1. **Comparison of execution time in minute between conventional codes, single node Hadoop and multi node Hadoop**

Different Methods	**Execution Time**		
Size of images file	**968MB**	**408MB**	**132MB**
Image mining in sequential codes	381 M	190.97 M	74 M
Image mining in single node Hadoop MapReduce framework	257 M	127.644 M	53 M
Image mining in multi node Hadoop MapReduce framework	215 M	105.514 M	31 M

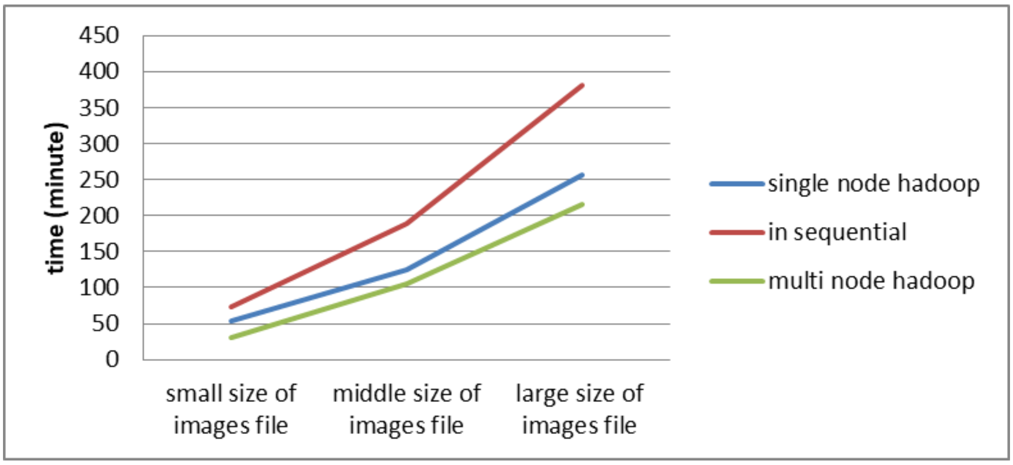

Figure 2. **Detection time on different size image files**

REFERENCES

[1] J. Kaur, K. Sachdeva, and G. Singh, "Image processing on multinode hadoop cluster," in *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 2017, pp. 21-26: IEEE.

[2] C. Ordonez and E. R. Omiecinski, "Image mining: A new approach for data mining," Georgia Institute of Technology1998.

[3] H. Min and Y. Shuangyuan, "Overview of image mining research," in *2010 5th International Conference on Computer Science & Education*, 2010, pp. 1868-1870: IEEE.

[4] P. Chouhan, M. J. I. J. o. A. R. i. C. Tiwari, and C. Engineering, "Feature Extraction Techniques for image retrieval using data mining and image processing techniques," vol. 5, no. 5, 2016.

[5] M. Maheshwari, S. Silakari, and M. Motwani, "Image clustering using color and texture," in *2009 First International Conference on Computational Intelligence, Communication Systems and Networks*, 2009, pp. 403-408: IEEE.

[6] A. Sleit *et al.*, "Image clustering using color, texture and shape features," vol. 5, no. 1, pp. 211-227, 2011.

[7] Y. Tang *et al.*, "OEHadoop: accelerate Hadoop applications by co-designing Hadoop with data center network," vol. 6, pp. 25849-25860, 2018.

[8] A. B. Patel, M. Birla, and U. Nair, "Addressing big data problem using Hadoop and Map Reduce," in *2012 Nirma University International Conference on Engineering (NUiCONE)*, 2012, pp. 1-5: IEEE.

[9] A. Madhuri, S. P. Praveen, D. L. S. Kumar, S. Sindhura, and S. S. J. A. o. t. R. S. f. C. B. Vellela, "Challenges and Issues of Data Analytics in Emerging Scenarios for Big Data, Cloud and Image Mining," pp. 412-423, 2021.

[10] E. ZAGAN, M. J. I. J. o. A. C. S. DANUBIANU, and Applications, "HADOOP: A Comparative Study between Single-Node and Multi-Node Cluster," vol. 12, no. 2, 2021.

[11] J. Ramsingh and V. J. A. S. C. Bhuvaneswari, "An integrated multi-node Hadoop framework to predict high-risk factors of Diabetes Mellitus using a Multilevel MapReduce based Fuzzy Classifier (MMR-FC) and Modified DBSCAN algorithm," vol. 108, p. 107423, 2021.

[12] S. Gugnani, D. Khanolkar, T. Bihany, and N. Khadilkar, "Rule based classification on a multi node scalable hadoop cluster," in *International Conference on Internet and Distributed Computing Systems*, 2014, pp. 174-183: Springer.

[13] A. O'Driscoll, J. Daugelaite, and R. D. J. J. o. b. i. Sleator, "'Big data', Hadoop and cloud computing in genomics," vol. 46, no. 5, pp. 774-781, 2013.

[14] S. M. Mahmoud, R. S. J. I. J. o. M. E. Habeeb, and C. Science, "Analysis of large set of images using MapReduce framework," vol. 11, no. 12, p. 47, 2019.

[15] R. J. C. E. Sudhir and I. Systems, "A survey on image mining techniques: theory and applications," vol. 2, no. 6, pp. 44-52, 2011.

[16] S. Silakari, M. Motwani, and M. J. a. p. a. Maheshwari, "Color image clustering using block truncation algorithm," 2009.

[17] S. Biswas, T.-h. Kim, and D. J. I. J. o. S. H. Bhattacharyya, "Features extraction and verification of signature image using clustering technique," vol. 4, no. 3, pp. 43-55, 2010.

[18] J. Zhang, G. Wu, X. Hu, S. Li, and S. Hao, "A parallel k-means clustering algorithm with mpi," in *2011 Fourth International Symposium on Parallel Architectures, Algorithms and Programming*, 2011, pp. 60-64: IEEE.

[19] Z. Lv, Y. Hu, H. Zhong, J. Wu, B. Li, and H. Zhao, "Parallel k-means clustering of remote sensing images based on mapreduce," in *International Conference on Web Information Systems and Mining*, 2010, pp. 162-170: Springer.

[20] M. Yang, H. Mei, and D. J. I. J. I. C. I. C. Huang, "An effective detection of satellite image via K-means clustering on Hadoop system," vol. 13, no. 3, pp. 1037-1046, 2017.

[21] W. K. Mutlag, S. K. Ali, Z. M. Aydam, and B. H. Taher, "Feature extraction methods: a review," in *Journal of Physics: Conference Series*, 2020, vol. 1591, no. 1, p. 012028: IOP Publishing.

[22] D. J. I. J. o. M. Ping Tian and U. Engineering, "A review on image feature extraction and representation techniques," vol. 8, no. 4, pp. 385-396, 2013.

[23] R. Roslan and N. Jamil, "Texture feature extraction using 2-D Gabor Filters," in *2012 International Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, 2012, pp. 173-178: IEEE.

[24] A. J. I. a. Humeau-Heurtier, "Texture feature extraction methods: A survey," vol. 7, pp. 8975-9000, 2019.

[25] V. K. Dehariya, S. K. Shrivastava, and R. Jain, "Clustering of image data set using k-means and fuzzy k-means algorithms," in *2010 International conference on computational intelligence and communication networks*, 2010, pp. 386-391: IEEE.