Implementation of Image Mining Technique Based Hadoop Map Reduce

Zahraa Azhar Muhammad Shamki
University of Kufa / Computer Science Department, AL-Najaf AL-Ashraf, Iraq

E-mail: Zahraa.alramahi@uokufa.edu.iq

Furkan Rabee
University of Kufa / Computer Science Department, AL-Najaf AL-Ashraf, Iraq

E-mail: furqan.rabee@uokufa.edu.iq

Received: 06th August, 2022; Accepted: 28th August, 2022; Published: 21st September, 2022

***Abstract*:** Image mining is a very important technique that is used to get information directly from images. In the era of big data and the fast growth in digital picture complexity Consequently, image mining technology now faces additional obstacles, such as a fair amount of time cannot be allotted for the analysis of such massive datasets using a personal computer or database process because there are limits to how much data can be stored and how much it can be computed. As a result, distributed computing is required for current image exclude collection mining. The Hadoop platform is an appropriate paradigm; because of its reliance on MapReduce functionalities. This paper gives an overview of the different image mining techniques that have been suggested in the past some of them are implemented using Hadoop map reduce.

Key Words: image mining, Hadoop, mapreduce, image processing, Bigdata.

1. INTRODUCTION

Due to developments in picture capturing and storage technologies, there has been a remarkable increase in extremely big and comprehensive image databases [1]. Satellite images, aerial shots, medical images, and digital photographs contribute to the daily production of vast amounts of image data. Users may learn about these images by using the right tools and approaches. As a result, humans are no longer able to process and evaluate such large amounts of images in order to extract the useful information or knowledge that may be obtained from each individual image. As a result, image mining has emerged as a useful tool for a wide range of applications[2]. Image mining is a crucial technology used to extract information directly from images. It is just a form of data mining that deals with images [3]. It is the process of getting hidden information from images, linking image data, and finding other patterns that aren't obvious. Machine learning, image processing, retrieval of image, mining the data, databases, computer vision, and artificial intelligence all benefit greatly from image mining, which is a multidisciplinary work. There has been an increase in the number of applications and approaches in the particular fields described above, But more research needs to be done on image mining. The challenges of image mining research and the growth of this field need to be looked into [4]. Researchers who work in data mining, information retrieval, and multimedia databases are starting to pay more attention to image mining as a result of its potential for uncovering useful visual patterns, which has the ability to push a number of different study disciplines to new frontiers[4]. Image mining is distinct from basic computer vision and image processing methods. Image mining focuses on the extraction of patterns from massive image collections. In contrast, image processing and computer vision algorithms focus on comprehending and/or gaining particular characteristics from an image. Images may conceal a wealth of information. In light of the ever-increasing volumes of image data, the need for image mining is tremendous, Image classification and clustering, as well as association rule mining, are some of the primary technologies used in image mining [5]. Massive size of the dataset and the intricate nature of the processes that are necessary for its processing, severe requirements for memory storage and computing performance are necessary. According to Google Trends, the search term "Hadoop" is the one that is most closely connected with "Big Data." Hadoop is, in fact, the most prominent technology that is associated with this issue [6]. Images and videos constitute the majority of these multimedia materials. In order to categorize, search, and tag pictures, therefore, image processing operations must execute in distributed systems.

There are distributed systems that enable saving and processing of massive amounts of data. Hadoop Distributed File System, or HDFS, is an open-source project that was made to do this

1.1. Image Mining Trends:

In image mining field, there are two primary trends [2]. The first trend focuses on domain-specific applications that extract the most pertinent image attributes in a format suited for data mining as shown in **(Hsu,et al., 2000)** [7] and **(Asanobu, KITAMOTO ,2001)** [8].

The second branch focuses on generic applications with the objective of producing visual patterns that can aid in Comprehending the relationship between high-level human perceptions and low-level image characteristics **(Ordonez, et al.,1999)** [9].

1.2. Apache Hadoop framework

Hadoop is a prominent framework for distributed computing that is used to manage large data mining challenges in data centers. 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 [10]. The processing component of Hadoop is referred to as MapReduce, while the storage component is referred to as the Hadoop Distributed File System (HDFS) [11].Hadoop framework is either single node or multi node Hadoop cluster [12]. In the next section of this paper gives an overview of the different image mining techniques that have been suggested before. The Methodology section explains process of image mining systems. In the last section, it comes to a conclusion.

2. LITRUTRE REVIEW

2.1. Feature Extraction and Image Clustering

Feature extraction is the main core in diagnosis, classification, clustering, recognition, and detection [13]. Image data is generally represented by characteristics like texture, edge, color, and shape. One of the key difficulties is how to represent this data, **(Foschi et al. ,2002)** [14] Presents A new feature extraction approach that uses a combination of color, edge, and texture properties to achieve excellent recall and precision. **Silakari et al. (2009)** Use “Color Moment” and “Block Truncation Coding (BTC)” to retrieve the image dataset's features, a structure was designed that emphasizes color as a distinguishing feature. Afterwards, the K-Means clustering algorithm is used to cluster the image collection into many groups. In the first step of evaluation, color moments for each of the three color components are calculated. Each color component gives a three-element feature vector: the mean, the standard deviation, and the skewness. So, for each image, a total of nine feature vectors are found. In the second part, use the Block Truncation Coding Algorithm on RH, RL, GH, GL, BH, and BL. So, 18 feature vectors are added up for a single image, where RH is made by taking the red part of all pixels in the image that are above the red average and RL is made by taking the red part of all pixels that are below the red average. In the same way, GH, GL, BH, and BL can be found. After that, the color moment can be calculated for RH, RL, GH, GL, BH, and BL. The results are good enough, and they show that the BTC algorithm works better than color moments. [15].

There are numbers of methods prevailing for Image Mining Techniques. **Dubey et.al. (2010)** in [16] includes the features of four techniques i.e. Color Histogram, Color moment, Texture, and Edge Histogram Descriptor then applying Image mining techniques depending on these features. The combination of the four techniques are used and the Euclidian distances are calculated of the every features are added and the averages are made. The results are quite good for most of the query images. Image clustering is often conducted early in the mining procedure[4]. Color, texture, and shape are the feature characteristics that have attracted the most attention for clustering. In general, any of the three might be employed alone or in combination [4]. There are several available clustering techniques. Once images have been clustered, a domain expert must review each cluster's images to name the abstract ideas represented by the cluster.

Biswas1 et al. (2010) proposed a strategy to extract characteristics from handwritten signature images and presented a way for recognizing handwritten signatures. That computed feature is used for verification. Here they used a clustering technique for verification. The current study investigates an image clustering method that is based on a k closest neighbors technique. This enables the process to manage clusters of varying sizes and shapes. The suggested system provides a relatively simple yet reliable solution to the problem of signature verification. The results of the experiments make it abundantly evident that this technique can, in fact, distinguish fake items from genuine ones [17].

Chang et al. (1999) Utilize clustering in an effort to detect unauthorized image duplication on the World Wide Web [18].

R.Venkata et al. (2012) Proposed system uses combinations of color feature and implemented on visual contents of an image such as color, shape, texture and spatial layouts. The suggested system chose 10000 image databases that had similar feature values and then extracted all of the images' features independently as R, G, and B values. These values are then used to get images from a big database. The proposed method is used to recover images from a huge database [19].

Zoghoul et al. (2017) generates effective strategies that may be applied to the process of extracting color image characteristics. These characteristics can function as a key or signature for a color image, allowing one to obtain or identify a color image. The approaches will be evaluated and put into practice, and an artificial neural network will be used to recognize colors in the images once the extracted characteristics have been sent to the network. The results of the experiments conducted with each of the new approaches will be compared in order to determine which one is the most effective in terms of producing the desired outcomes[20].

A.Kannan et al (2010) the primary purpose of image mining is to eliminate data loss and recover potentially valuable information relevant to human needs. Existing Content Based Image Retrieval (CBIR) systems are many in the current context. However, this specific method will be used in Medical transcription in an efficient manner not only based on the image's content, but also on the query image, for the purpose of comparing some common diseases that afflict human bodies at an earlier age. In this system, a new cluster-based image retrieval approach is also used to minimize the searching time and space. In addition, the RGB components of the color images are categorized in different dimensions to form Red, Blue, and Green image clusters [21].

Texture characteristics are retrieved from the full image and then utilized to determine the degree of similarity between two images.

D.Zhang et al. (2000) present a method for image retrieval based on the Gabor filter. Calculating the mean and standard deviation of a Gabor-filtered image yields information about the texture. A circular shift of the feature components is used to achieve rotation normalization so that all images have the same dominant direction. Textured and natural images are used in the indexing and retrieval of images[22].

Clustering is one of the most popular techniques in image mining It is common in a variety of fields, including image segmentation, bioinformatics, pattern recognition, and statistics, etc. K-means is the most common and easiest clustering method due to its simple implementation, efficiency, and empirical performance. Nonetheless, real-world applications generate enormous amounts of data, making the efficient management of this data in an essential mining activity a tough and major challenge. In addition, as a message-passing programming model, MPI (Message Passing Interface) offers excellent speed, scalability, and portability. In response, MKmeans, a parallel K-means clustering technique with MPI, was developed, is proposed in [23] by **J.Zhang et al (2011).** 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 [24] we are also doing image mining technique that is image clustering in Hadoop mapreduce. Images in our data collection have pixels that migrate to an HDFS file when viewed. Features are defined as values obtained from a specific color layer of Red, Green, and Blue. After extracting color and texture features for each image, the K-Means clustering technique is used to group the photos into clusters based on their similarity using Hadoop mapreduce. We may compare the execution time of sequential codes and the Hadoop system based on the experiment and findings. Using the Parallel K-Means technique, it was shown to be 24% quicker than the standard clustering algorithm.

2.2. RS Image clustering using Hadoop

RS (remote sensing) images, whose characteristics need to be identified or grouped to offer research basis, are some of the key studies in the subject of deforestation, climate change, ecology, and land surface temperature. For RS images, K-Means clustering is an essential approach for providing an immediate overview of the elements. For PCs, processing a high number of RS images is hindered by the limited hardware resources and the tolerance for time-consuming tasks. Parallel computing and distributed systems are without question the best options. Unlike traditional ways, Z. Lv et al. (2011) in [25] They attempt to parallelize K-mean algorithm using Hadoop to cluster the objects inside RS images , Hadoop implements the MapReduce programming model to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Images are converted into text files to fit the framework. Each line in form: (file_id, pixel_id, r, g, b) represents a pixel. Configure, map, and close are the three primary functions of a Hadoop job. The user's description file specifies starting centroids for the task, which are read in the configure function. Most clustering computations are handled by the map function. The final centroids are saved to a file when the close function is used (called centroid file). When the job is committed to Hadoop, the input pixels of one map task are just a subset of the total number of pixels. Several centroid files need to be combined before the next step can be taken. Mean of each cluster centroid is used as the new centroid here. There is no additional work performed by the reduce functions in each task other than the writing of the end output to files (called result file). That's how cluster values for each pixel's cluster come to be included in the results files. Their work demonstrates that the outcomes are satisfactory and may potentially inspire new techniques to addressing comparable difficulties in the realm of remote sensing applications.

due to the huge amount of satellite photos in the era of Big Data, satellite image detection is typically challenging, therefore fast processing speed has become a necessity for certain applications, such as rapid catastrophe warning reaction M.Yang et al. (2017) also parallelize K-Means clustering on the Hadoop system such as in [25] but to detect satellite images effectively instead of RS images and some differentiation in images sizes [26]. They design the effective K-Means algorithm using MapReduce programming model and Hadoop distributed file system to clustering objects inside satellite images. Two main operations in MapReduce: Map and Reduce, are realized to give an efficient implementation. The results show a fast detection speed and good scaleup while keeping accuracy both in training and testing.

Mohamed H. Almeer in [27] On the Hadoop MapReduce parallel platform, remote sensing-related image processing algorithms were examined and implemented employing an experimental 112-core high-performance cloud computing machine housed at the Environmental Studies Center at the University of Qatar. In this examination, they presented a case study for simultaneously processing TIF-formatted remote sensing photos with the Hadoop MapReduce framework. When applied to remote sensing photos, the experimental findings demonstrate that conventional image processing methods may be parallelized with acceptable execution speeds. A vast number of pictures cannot be processed sequentially in an effective manner. Despite its original intended for text processing, Hadoop MapReduce applied on a parallel cluster has demonstrated success in processing vast quantities of TIF pictures. When it comes to custom processing in the Hadoop MapReduce architecture, researchers have been looking for an efficient programming solution. Hadoop was used to assess the processing speed of massive Landsat image archives. In all, eight image analysis techniques were employed: Sobel filtering, image scaling, image format conversion, auto contrasting, image sharpening, picture embedding, text embedding, and image quality inspection Results show MapReduce has the capacity to handle large-scale remotely sensed pictures and solve increasingly complicated geographical challenges.

Golpayegani and Halem [28] Hadoop is being used to process satellite images. They provide a text-based parallel content search. After being received by satellites, each image is therefore tagged with textual information. Similarly Krishna et al. [29] Using Hadoop distributed file system for storage and Map/Reduce paradigm for processing web-crawled pictures, this article shows how to build and construct a hybrid distributed architecture. This design leverages the capability of Hadoop's Map/Reduce framework with stand-alone crawler nodes to get relevant material from the web when the process has to be parallelized. An evaluation of the system's ability to store and analyse billions of photos in a few hours using real-world online data shows that. To get started, all you need is a collection of picture URLs and any associated metadata. The system as a whole is designed to process incremental data quickly. The incremental data is generated by comparing any incoming feed to the HDFS dataset. New URLs are scheduled to be scanned by the content fetcher module, and any changes to the metadata of previously crawled pictures are immediately updated on HDFS. A high number of documents must be retrieved, processed, and stored in order for this architecture to be appropriate for multimedia applications.

2.3. Fingerprint Image mining using Hadoop

S. Mahmoud, R. Habeeb, and C. Science (2019) [30] was used image mining and MapReduce programming approach to analyze a huge amount of fingerprint images that cannot ordinarily be handled owing to a limited physical memory in order to identify the images' characteristics simultaneously. The images are initially retained in image data storage in order to be preprocessed and to extract the features for each user's biometric characteristic before being stored in a database. Mapper and reducer routines are utilized to cope with this enormous batch of fingerprint pictures. In order to extract characteristics from several photos, the map function discovers the morphological thinning, ridges terminating, and bifurcation while the reduction function gathers them. The results show a considerable reduction in the amount of time it takes to construct a feature vector for each image.

2.4. Medical image mining using Hadoop

Image mining although being a young study field, its progress shows significant promise, and its application has already spread to other sectors with positive outcomes. Mining on medical pictures is the process of acquiring important knowledge and models, which can then be applied to the discovery of anomalous circumstances that are inconsistent with the most prevalent models. This can serve as a reference and assist physicians in diagnosing diseases.

In the last 20 years, both the number of images and the number of dimensions they have grown by a lot. Recent advances in image processing and machine learning make it possible to help clinicians find and describe important events in long series of images. But the process of extracting complex features from large sets of 3D/4D images, training machine learning algorithms, and optimizing the whole system globally is very time-consuming, storage-intensive, and network-bandwidth-intensive [31] Large-scale picture description and analysis have recently been performed using the MapReduce framework, a distributed computing platform. **Markonis, et.al (2012) in [31]**, MapReduce was utilized to speed up and enable three large-scale medical image processing use cases: (i) parameter optimization for lung texture classification using support vector machines (SVM), (ii) content–based medical image indexing, and (iii) three–dimensional directional wavelet analysis for solid texture classification.

S.Jyoti and G. Pradeepani (2016) In [32], using “Hadoop MapReduce Framework”, offer a method for evaluating medical images. As a way to deal with massive amounts of medical picture data, they propose a Hadoop MapReduce-based system for retrieving and analyzing images. A prototype system with a master node and three subordinate nodes is under consideration. A MapReduce job is generated when the master node receives a query picture from the user and uses it to create the query image's feature vector. Afterwards, the master node will construct a new MapReduce job to match the feature vector of the query picture with the feature vectors in the HDFS feature library. Using the last MapReduce task's output, a third MapReduce task will be run to gather all of the necessary photos. Parallelism enables the Hadoop-based medical image retrieval system to locate pictures significantly quicker than previous image retrieval systems.

2.5. Processing high-resolution satellite image data using Hadoop

Using parallel and distributed processing capabilities of the Hadoop MapReduce Framework, complicated image processing algorithms requiring more computational power and larger inputs may be executed effectively. Hadoop MapReduce is a scalable approach for handling petabytes of data with enhanced fault tolerance and data parallelism. **(Roshan Rajak,et.al)** in [33] They provide a MapReduce framework for processing satellite remote sensing data simultaneously using Hadoop and storing the results in HBase. By employing Hadoop, enterprises can split their workload over several clusters and leverage the pooled processing power of commodity hardware, as seen by the speed and performance benefits. This study examines the efficacy of Hadoop cluster architecture for processing high-resolution satellite image data by optimizing HDFS storage. This study addresses fundamental challenges in remote sensing, such as the necessity for intense processing. Although they analyzed satellite remote sensing data for their implementation, the framework may also be used to manage photos from medical networks, social media, and industries. Even for complicated image processing algorithms, it was demonstrated that a minimum of 7X performance could be reached with only a four-node cluster. The results demonstrate that the suggested map-reduce architecture significantly accelerates the processing of massive image data.

2.6. Grid Computing and Hadoop

Grid computing is a networking technology that lets people use the processing power of computers all over the world at the same time. Grid networks have been used in the medical field for a few years now. Images from medical tests like X-rays, MRI scans, CT scans, etc. need to be shared between the doctors and technicians who look at the images and figure out what's wrong. If there aren't any good ways to store these images, there could be a bottleneck in the system because people are trying to get in at different times. One new trend in medical image retrieval is the use of Hadoop framework for big data analytics[34], **R. Kingsy, et al. (2014)** in [34] Using the Apache Hadoop framework, that is compatible with a variety of image formats and may be used to store, distribute, and retrieve photos in hospitals as shown in Fig.1 [34]. Using Hadoop improves several performance parameters, including reliability, accuracy, interoperability, privacy, and security. User authentication is also done, and patient privacy is kept safe. The texture-based Content Based Image Retrieval (CBIR) method is used in this work to quickly find images. Experiment findings indicate that Hadoop-performance CBIRs increase significantly when the database contains a large number of photos.

2.7. Biomedical imaging application and Hadoop

K. Neshatpour et al. (2016) offers a way to accelerate the processing of enormous biomedical imaging applications using FPGA platform. For acceleration, the Laplacian filtering of images to find edges and the K-means clustering of biomedical images have been looked at.

The two algorithms' MapReduce implementation has been built, and the computationally heavy jobs have been translated to the FPGA. The results show that Hadoop Map might be able to speed up the kernel for large image data sets. The framework and technique may be applied to speed other computationally expensive biomedical image processing applications that deal with vast amounts of imaging data[35]

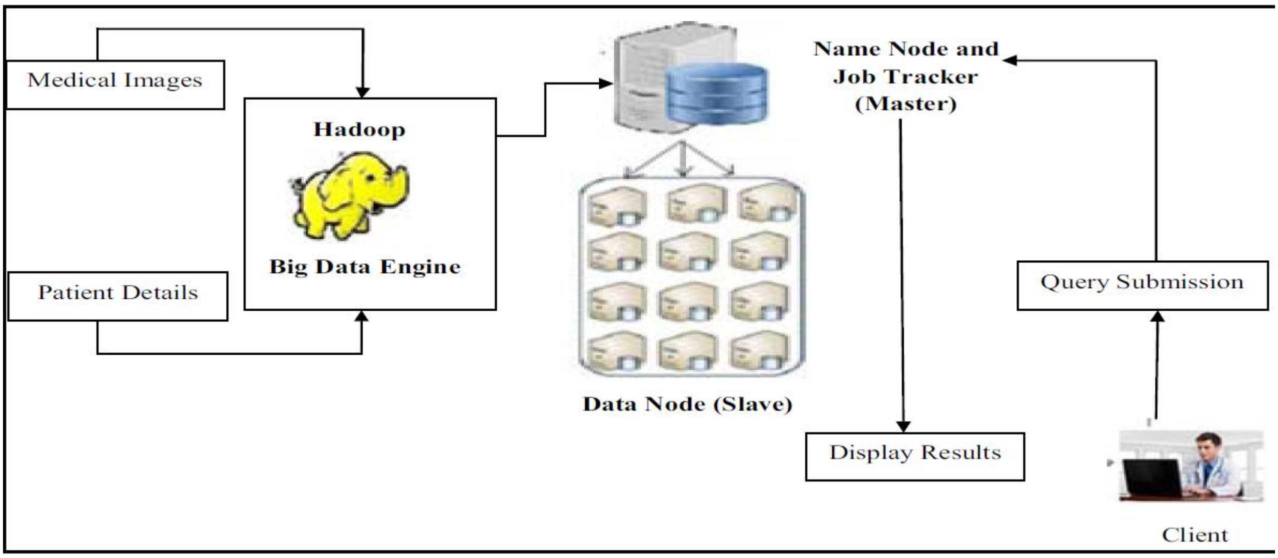

Fig 1. Data Grid architecture

2.8. Face Detection and Hadoop

Applications of face detection are commonly used for finding, labeling, and categorizing individuals inside vast picture datasets. This sort of application demands the processing of a huge number of relatively small-sized photos. On the other hand, the Hadoop Distributed File System (HDFS) was created to store and process big files. A large number of small-sized pictures slows down HDFS by increasing overall job initiation time, task scheduling overhead, and file system manager memory consumption (Namenode).The study in [36] by **Demir and Sayar (2014)** describes two methods for improving the speed of HDFS while processing small image files. These include (1) merging the photos into a single huge file and (2) integrating many photographs for a single purpose without merging. In order to develop these strategies, Additionally, they provide Hadoop file formats and record production techniques (for reading picture information). In this method, they apply a sophisticated face recognition methodology by evaluating pixel information in the photographs and employing pictures in their natural forms. The picture was severed at a probable face region.

Dong et al. offer two strategies given in [37] and [38] for HDFS problem. In [37] They provide techniques for merging and perfecting files that are structurally linked, as well as ways for grouping and perfecting files that are logically connected. In this approach, files are classified according to their logical or structural characteristics. In [38], They take a similar approach to the same issue. They describe a two-level technique for improving the efficiency of accessing tiny files, using PowerPoint files as an illustration [36].

Kocakulak and Temizel in [39] Hadoop and MapReduce were used to do ballistic image analysis, which involves integrating a massive photo collection with an unknown image. This method relied on a correlation technique that demanded substantial computing resources. It has been proved that processing time is drastically reduced when 14 computational nodes are utilized in a cluster configuration.

2.9. Image processing and Hadoop

Karan,Jaideep and Gursimran [40] In this work, Image Processing was conducted on a MultiNode Hadoop Cluster with amazing results. Around 100MB of photographs were gathered using cameras, the Internet, and mobile devices. Using the Hadoop framework and HIPI, it was possible to extract the Hadoop Image Bundle's metadata properties. The whole duration of processing was less than one minute and thirteen seconds. Using Hadoop and HIPI onMultiNode Hadoop Cluster, it was observed that metadata information can be collected from unstructured data with efficiency. Hadoop technology processes this data far faster than previous methods. This demonstrates that Hadoop will always be superior to traditional approaches.

3. METHODOLOGY

Observing existing image mining systems, the total procedure may be broken down into the following components:

A. Data preprocess:

Eliminate noisy data in huge image databases.

B. Extracting feature vectors:

The practice of using image processing technology to extract feature vectors linked to a certain activity is referred to as "extracting feature vectors."

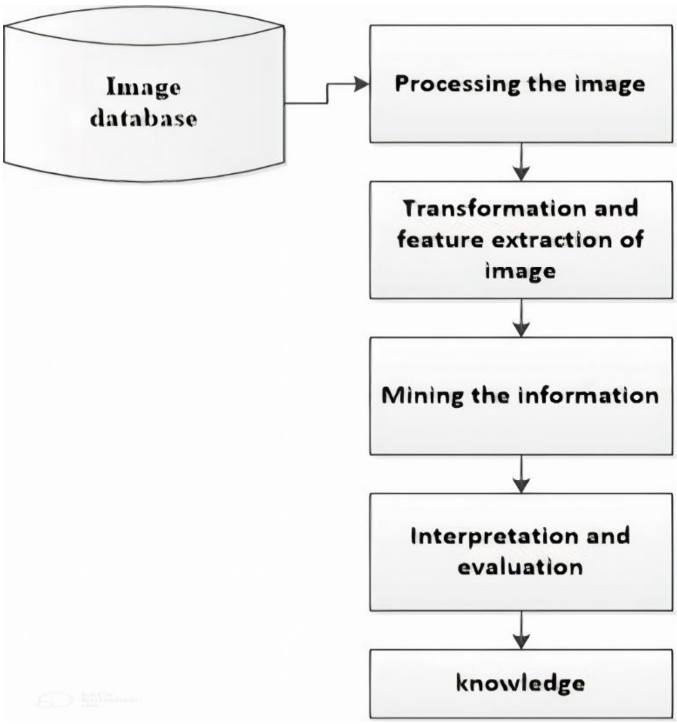

Fig 2. Image mining process

C. Extraction of information from vectors and acquisition of advanced knowledge:

On feature vectors, a variety of techniques, such as object identification, image indexing and retrieval, image classification and clustering, and neural networks, are utilized in order to mine and acquire useful and hidden high-level knowledge, which is then evaluated and explained [5]. Fig.2 shows the Typical Image Mining Process[41].

4. CONCLUTION

In this paper provides an overview of several image mining strategies offered by researchers in the past. The objective of mining is to generate all essential patterns without knowing them beforehand. Images can conceal crucial information; nevertheless, few researches examine image data mining. Image mining is only an expansion of data mining within the realm of image processing. Some of the studies above are done in Hadoop map reduce to overcome problems of limit storage and time of processing. All previous research results such as extract the Image Bundle's metadata properties using Hadoop and HIPI on MultiNode Hadoop Cluster, can be collected in less than one minute and thirteen seconds and Hadoop proved its efficiency in face detection and biomedical imaging application and more other that mentioned above. These results proved that the Hadoop can process large amounts of images in less time than normal methods, which led to its use in many engineering, medical and RS image applications to make more effort in less time.

References

[1] Zaiane, O.R., et al. Mining multimedia data. in CASCON. 1998.
[2] Ennouni, A., et al. A review on image mining. in 2017 Intelligent Systems and Computer Vision (ISCV). 2017. IEEE.
[3] Sudhir, R.J.C.E. and I. Systems, A survey on image mining techniques: theory and applications. 2011. 2(6): p. 44-52.
[4] Hsu, W., M.L. Lee, and J.J.J.o.i.i.s. Zhang, Image mining: Trends and developments. 2002. 19(1): p. 7-23.
[5] Min, H. and Y. Shuangyuan. Overview of image mining research. in 2010 5th International Conference on Computer Science & Education. 2010. IEEE.
[6] De Mauro, A., M. Greco, and M. Grimaldi. What is big data? A consensual definition and a review of key research topics. in AIP conference proceedings. 2015. American Institute of Physics.
[7] Hsu, W., M.L. Lee, and K.G. Goh. Image mining in IRIS: Integrated retinal information system. in Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 2000.
[8] Asanobu, K. Data mining for typhoon image collection. in Proceedings of the second international conference on multimedia data mining. 2001. Citeseer.
[9] Ordonez, C. and E. Omiecinski. Discovering association rules based on image content. in Proceedings IEEE Forum on Research and Technology Advances in Digital Libraries. 1999. IEEE.
[10] Tang, Y., et al., OEHadoop: accelerate Hadoop applications by co-designing Hadoop with data center network. 2018. 6: p. 25849-25860.
[11] 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.
[12] 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.
[13] 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.
[14] Foschi, P.G., et al. Feature Extraction for Image Mining. in Multimedia information systems. 2002.
[15] Silakari, S., M. Motwani, and M.J.a.p.a. Maheshwari, Color image clustering using block truncation algorithm. 2009.
[16] Dubey, R.S., et al., Multi feature content based image retrieval. 2010. 2(6): p. 2145-2149.
[17] Biswas, S., T.-h. Kim, and D.J.I.J.o.S.H. Bhattacharyya, Features extraction and verification of signature image using clustering technique. 2010. 4(3): p. 43-55.
[18] Chang, E.Y., et al. Searching near-replicas of images via clustering. in Multimedia Storage and Archiving Systems IV. 1999. SPIE.
[19] Chary, R., D.R. Lakshmi, and K.J.a.p.a. Sunitha, Feature extraction methods for color image similarity. 2012.
[20] Al Zudool, M.A., S. Khawatreh, and Z.A. Alqadi, Efficient Methods used to Extract Color Image Features. 2017, IJCSMC.

[21] Kannan, A., V. Mohan, and N.J.a.p.a. Anbazhagan, An effective method of image retrieval using image mining techniques. 2010.
[22] Zhang, D., et al., Content-based image retrieval using Gabor texture features. 2000. 3656: p. 13âĂŞ15.
[23] Zhang, J., et al. A parallel k-means clustering algorithm with mpi. in 2011 Fourth International Symposium on Parallel Architectures, Algorithms and Programming. 2011. IEEE.
[24] Z. A. M. . Shamki and F. . Rabee, “Image mining technique using Hadoop map reduce over distributed multi-node computers connections”, AJEST, vol. 1, no. 2, pp. 18–24, Jul. 2022.
[25] Lv, Z., et al. Parallel k-means clustering of remote sensing images based on mapreduce. in International Conference on Web Information Systems and Mining. 2010. Springer.
[26] Yang, M., 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. 2017. 13(3): p. 1037-1046.
[27] Almeer, M.H.J.J.o.E.T.i.C. and I. Sciences, Cloud hadoop map reduce for remote sensing image analysis. 2012. 3(4): p. 637-644.
[28] Golpayegani, N. and M. Halem. Cloud computing for satellite data processing on high end compute clusters. in 2009 IEEE International Conference on Cloud Computing. 2009. IEEE.
[29] Krishna, M., et al. Implementation and performance evaluation of a hybrid distributed system for storing and processing images from the web. in 2010 IEEE Second International Conference on Cloud Computing Technology and Science. 2010. IEEE.
[30] Mahmoud, S.M., R.S.J.I.J.o.M.E. Habeeb, and C. Science, Analysis of large set of images using MapReduce framework. 2019. 11(12): p. 47.
[31] Markonis, D., et al. Using MapReduce for large-scale medical image analysis. in 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology. 2012. IEEE.
[32] Patil, J.S., D.G.J.I.J.o.s. Predeepini, and Technology, Two dimensional medical images diagnosis using mapreduce. 2016. 9(17).
[33] Rajak, R., et al. High resolution satellite image processing using Hadoop framework. in 2015 IEEE international conference on cloud computing in emerging markets (CCEM). 2015. IEEE.
[34] Grace, R.K., R. Manimegalai, and S.S. Kumar. Medical image retrieval system in grid using hadoop framework. in 2014 international conference on computational science and computational intelligence. 2014. IEEE.
[35] Neshatpour, K., et al. Big biomedical image processing hardware acceleration: A case study for K-means and image filtering. in 2016 IEEE International Symposium on Circuits and Systems (ISCAS). 2016. IEEE.
[36] Sayar, A.J.I.J.o.C.C. and Control, Hadoop optimization for massive image processing: case study face detection. 2014. 9(6): p. 664-671.
[37] Dong, B., et al., An optimized approach for storing and accessing small files on cloud storage. 2012. 35(6): p. 1847-1862.
[38] Dong, B., et al. A novel approach to improving the efficiency of storing and accessing small files on hadoop: a case study by powerpoint files. in 2010 IEEE International Conference on Services Computing. 2010. IEEE.
[39] Kocakulak, H. and T.T. Temizel. A Hadoop solution for ballistic image analysis and recognition. in 2011 International Conference on High Performance Computing & Simulation. 2011. IEEE.
[40] Kaur, J., 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. IEEE.
[41] Chouhan, P., 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. 2016. 5(5).

Authors' Profiles

Zahraa-Shamki is an MSc student in the Computer Science Department, Faculty of Computer Science and Mathematics, at the University of Kufa. She got BSc in Computer Science Department, Faculty of Computer Science and Mathematics, University of Kufa in 2015.

Furkan-Rabee is a staff member in the Computer Science Department, Faculty of Computer Science and Mathematics, at the University of Kufa. He got BSc and MSc in Computer Engineering from AL- Nahrian University in 2000 and 2008. He got Ph.D. in Computer Science and IT from the School of Computer Science and Engineering, UESTC, Chengdu, China 2015. The research interests include real-time scheduling algorithms, real-time locking protocols, Operating Systems, Parallel Processing, Distributed System, Computer Network, IoT, Mobile Computing, Cloud Computing, and Smart Cities.