Precise & Proficient Image Mining Using Hierarchical K-Means Algorithm

Parag Dhonde *, Prof. C. M. Raut **

* M.E. (Computer), Datta Meghe College of Engineering, Navi Mumbai – 400708, India
parag.dhonde@gmail.com
** Asst. Prof., Dept. of Computer Engineering, Datta Meghe College of Engineering, Navi Mumbai – 400708, India
cmr.cm.dmce@gmail.com

Abstract- Increasing use of World Wide Web and communication channels like mobile networking has increased the number of images used throughout the world. Continuing advancements in both hardware and software coupled with higher image processing and image vision tools, have made it possible to store huge amount of images. This increase in number of images and image databases has necessitated the need for image mining. The extension of data mining into the image domain is known as image mining. It is an interdisciplinary endeavour that draws upon expertise in computer vision, image processing, image retrieval, data mining, machine learning, database, and artificial intelligence. Rather than the development of many algorithms and applications in various research fields, image mining is still rarely untouched area. It mainly comprises of faster image retrieval and quality of the retrieved image. The extraction of implicit knowledge, image data relationship, and similar type of patterns may be the possible candidature to speed up the process of image retrieval. The proper combination and parameterization of these attributes can help out to retrieve better images at short point in time. This paper incorporates such two algorithms namely- hierarchical and k-means to have a good quality image retrieval at efficient pace.

Index Terms- Image processing, Image Mining, Image Retrieval, K-mean, Clustering

I. INTRODUCTION

A vast amount of image data is generated in daily life as image data plays vital role in every aspect of the systems like business, hospitals, engineering and so on. Image mining is the study of new technologies which helps to analysis and interpretation of the images and thus helps in development. Image retrieval is at the heart of this complete process. It is still at the experimental stage and growing field of research. Image retrieval is the process of browsing, searching and retrieving images from a large database of digital images. The collection of images in the web are growing larger and becoming more diverse. Retrieving images from such large collections is a challenging problem. One of the main problems is, it is hard to locate a required image in a huge and varied collection. On the other hand it is very much possible to identify a desired image from a small set simply by browsing, but the much more effective techniques are needed with collections contain thousands of items. To search for images, a user may provide query terms such as keyword, image file/link, or click on some image, and the system will return images "similar" to the query. The similarity used for search criteria could be meta tags, color distribution in images, region/shape attributes, etc. Unfortunately, image retrieval systems have not kept pace with the collections they are searching. The shortcomings of these systems are due both to the image representations they use and to their methods of accessing those representations to find images. The problems of image retrieval are becoming widely recognized, and the search for solutions an increasingly active area for research and development. The number of features required to represent an image can be very huge[1]. This paper proposes a novel image retrieval system which focuses on features of input query image as compared to the features of whole database image. Here hierarchical and k-means clustering is applied on database images so that query image's low level features as texture and shape are compared to only clustered images features rather than whole database image's features to improve speed, accuracy and efficiency. Thus, accuracy and time have been analyzed due to clustering in database images for retrieval.

II. LITERATURE SURVEY

Chary et al. [2] described the retrieval of images within a large image collection based on color projections and different mathematical approaches which are introduced and applied for retrieval of images. Images are sub grouping using threshold values, they considered R, G, B color combinations for retrieval of, which are implemented and results are included, and through results it is observed that it obtaining efficient results comparatively to the previous one and existing. This method provides the best solution in large image set compared with total of 10000 images with different categories. S. Balan and T. Devi. [3] explains that the retrieval process represents a visual query to the system and extracts the images based on the user request such mechanism referred to as query-by-example and used to compare some similarity metrics to compare query and target images. The greater demand for retrieval and management tools for visual data and visual information is a more capable medium of conveying ideas and is more closely related to human percept ion of the real world. Kreftin et al. [4] proposed the grid integration of medical image processing applications as grid workflows, where the workflow manager is responsible for the execution of all tasks related to grid communication and the developer is responsible for setting the access rights on his code and defining

the workflow manager what to do with it. Coarse-grained parallelization of processing steps can be applied in order to achieve runtime reduction. Kooper et al. [5] presented a novel solution for reconstructing 3-D medical volumes proposing the WS implementation as an additional layer to a dataflow framework. Olabarriaga et al. [6] developed a set of medical image analysis tools and described the requirements in order to integrate the existing systems (frameworks) with them. Stanchev [7], using image mining in image retrieval, described a new method for image retrieval using high level semantic features. The method includes the extraction of texture characteristics, low-level color, shape and various high level semantic features using the rules of the fuzzy logic helps in image mining technique. Herbert Daschiel and Mihai Datcu [8] explains the concept of Knowledge-driven content-based information mining system created to elaborate large amount of remote sensing image data. The system contain of intensive online and offline interface. The offline section include the extraction of ancient image features(such as facial expression, texture), their firmness, and data lessening, the generation of a entirely unsupervised image content-index, and the ingestion of the catalogue entry in the database management system. R. Brown, B. Pham. [9] described in detail a broad hierarchical image classifier approach and illustrated with which it can trained to find objects using support vector machine concept. In this approach speed and time complexity of algorithm is not discussed. Aura Conci. , Everest Mathias, M.Castro [10] proposed a framework for mining images by colour content. Their framework provides the possibility of use 5 distance function for evaluation of similarity among images and 2 types of quantization.

III. SYSTEM ANALYSIS

Main issues in analyzing images are the effective identification of features and another one is extracting them. For centuries, most of the images retrieval is text-based which means searching is based on those keyword and text generated by human's creation. Unfortunately, image retrieval systems have not kept pace with the collections they are searching. Most of the image retrieval systems present today are text-based, in which images are manually annotated by text-based keywords and when we query by a keyword, instead of looking into the contents of the image, this system matches the query to the keywords present in the database. The text-based image retrieval systems only concern about the text described by humans, instead of looking into the content of images. Data Clustering is often took as a step for speeding-up image retrieval and improving accuracy especially in large database. In general, data clustering algorithms can be divided into two types: Hierarchical Clustering Algorithms and Non-hierarchical Clustering Algorithms. However, Hierarchical Clustering is best suitable for clustering small quantities of data. Non- hierarchical is suggested to cluster large quantities of data. The most adapted method for non-hierarchical clustering is the K-Means clustering algorithm. This obviously suggests reducing the quantities of images prearranged to hierarchical clustering to have a faster retrieval. We can combine the image attributes such as color, texture, and shape in order to have a small set of data images. For this small set of data images, hierarchical cluster will retrieve images faster. These images later combined with K-Means clustering algorithm will produce better and accurate images. Thus, we propose an integrated approach to combine merits of the hierarchical clustering and k-means and discard disadvantages mentioned above of hierarchical and k-means algorithm. This method is different from the existing methods which first includes hierarchical clustering to confirm the number of clusters with location and then run the K-means clustering. The idea of the concept is to cluster first half data by hierarchical clustering and rest half by K-means in one round.

IV. SYSTEM DESIGN

The proposed system can be portrayed in the fig 4.1.Initially we create the image database. Then we extract the query image depend on the combination of color, texture, and shape of selected dataset. Now, the clustered images from the hierarchical clustering are applied to the k-means algorithm which takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high.

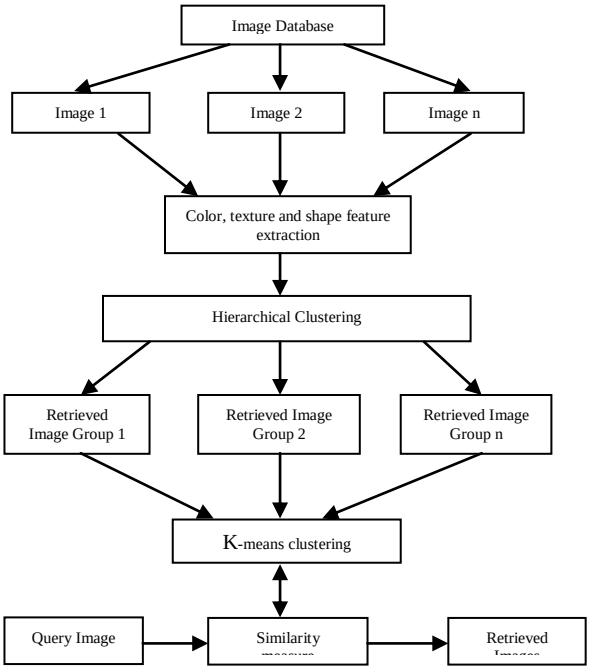

Figure 1: Image Mining System

An object is assigned to the cluster to which it is the most similar one. This object assignment is based on the distance between the object and the center it's closest to. It then computes the new centroid and in this way each center finds the centroid of its own points. This process iterates until the criterion function converges. Thus, the retrieval will be very accurate with the hierarchical and K-Means clustering. It leads to the better performance than by using individual algorithmic methods.

V. IMPLEMENTATION

To implement the above proposed system, we have used MATLAB as a tool. MATLAB is a state-of-the-art mathematical software package, which is used extensively in both academia

and industry. It is an interactive program for numerical computation and data visualization, which along with its programming capabilities provides a very useful tool for almost all areas of science and engineering. MATLAB makes use of highly respected algorithms and hence we can be condensing about our results. Powerful operations can be performed using just one or two commands. We can build up your own set of functions for a particular application. MATLAB is a high-performance language for technical computing. It integrates calculation, visualization, and coding in an easy-to-use environment where tribulations and solutions are expressed in well-known mathematical notation.

VI. RESULT & DISCUSSION

For this approach we have used performance parameter as- time, relevant images, accuracy, Recall and Precision.

A. Accuracy

Accuracy of an image retrieval task is defined as the ratio of the number of relevant images retrieved to the total number of images retrieved expressed in percentage.

$$Accuracy = \frac{number\ of\ relevant\ images}{total\ number\ of\ images\ retrieved} \times 100 \quad (1)$$

Where, total number of images retrieved = number of relevant images + number of irrelevant images

An assumption is made to calculate the accuracy values using the first 50 relevant image results for uniformity and simplicity of calculations. Accuracy is a vital parameter for evaluation as it is a direct measurement of the quality and user satisfaction of the image retrieval process. It is summarised in Table 1

B. Recall

Recall measures the capability of the system to retrieve all models that are relevant, while precision measures the capability of the system to retrieve only models that are relevant. Recall can be calculated as:

$$Recall = \frac{number\ of\ relevant\ images\ reterived}{total\ number\ of\ relevant\ images} \quad (2)$$

C. Precision

It can be calculated as:

$$Precision = \frac{number\ of\ revelant\ images\ retrieved}{total\ number\ of\ images\ retrieved} \quad (3)$$

The first test is based on the color dominant and texture features approach. In this test, the 4 image classes are taken and we have calculated the precision and the recall for these different image classes in Table 2.

TABLE I
RESULTS OBTAINED BASED ON COLOR, TEXTURE, SHAPE AND COMBINED APPROACH

Image Class	Time (Seconds)	Image Received	Relevant Images	Accuracy
Classification based on color	1.8544	11	3	27%
Classification based on shape	1.8274	11	2	18%
Classification based on texture	1.8385	11	2	18%
Classification based on color, shape and texture	1.8191	11	9	81.81%

TABLE III
RECALL AND PRECISION FOR 4 IMAGE CLASS

TEST	IMAGE 1	IMAGE 2	IMAGE 3	IMAGE 4
Class	TajMahal	Great Wall of China	Wall	Flowers
Query Image Type	Jpg	Jpg	Jpg	Jpg
Matched Relevant Images	6	6	7	5
Precision	0.75	0.66	0.77	0.62
Recall	0.86	0.75	0.87	0.71

The comparison between the individual approach and combined approach using the parameters is shown as below:

TABLE III
COMPARISION BETWEEN INDIVIDUAL APPROACH AND COMBINED APPAROACH

Parameters	Individual Approach	Combined Approach
MeanTime(sec)	1.8336	1.8191
Mean Accuracy(%)	21%	81.81%

So, it can be clearly visible that the combined approach gives a much more balanced performance in terms of all two parameters. While keeping the mean time taken for image retrieval below 2 seconds, we can achieve an accuracy of 81.81%.

VII. CONCLUSION

As day by day use of images, pictures is increasing it makes image database very large, thus to get the desired image it takes lot of time. So the difficulty is not the technology it basically very large database. To reduce such large database for desired images this paper suggest hierarchical algorithm. Hierarchical algorithm gives the best result for small dataset. Pre processing of image based on color, texture and shape help to reduce the dataset this lead to proficient retrieval at the same time to ensure the precise image retrieval we incorporates k-means algorithm. Thus hierarchical and k means supports the proficient and precise image mining respectively. This combinatory approach has 81.81% proficiency. Thus using hierarchical and K-Means techniques together not only facilitates the user not to overlook the image he may require but also to obtain accurate favored image results proficiently.

REFERENCES

[1] Khodaskar. A.A, Ladhake. S.A.:Image Mining: An Overview of Current Research, IEEE April 2014.

[2] R.VenkataRamana Chary, Dr.D.Rajya Lakshmi and Dr. K.V.N Sunitha "feature extraction methods for color image similarity",Advanced Computing: An International Journal (ACIJ), Vol.3, No.2, March 2012.

[3] S.Balan and T.Devi,"Design and Development of an Algorithm for Image Clustering In Textile Image Retrieval Using Color Descriptors", International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.3, June 2012.

[4] D. Kreftin, M. Vossberg, A. Hoheisel, and T. Tolxdorff, "Simplified implementation of medical image processing algorithms into a grid using a workflow management system," J. Future Gener. Comput. Syst., vol. 26, no. 4, pp. 681–684, Apr. 2010.

[5] R. Kooper, A. Shirk, S.-C. Lee, A. Lin, R. Folberg, and P. Bajcsy, "3D medical volume reconstruction using web services," Comput. Biol. Med., vol. 38, no. 4, pp. 490–500, Apr. 2008.

[6] Aura Conci and Everest Mathias M. M. Castro, "Image mining by content", Expert Systems with Applications, Vol. 23, no. 4, pp. 377-383, April 2002.

AUTHORS

First Author – Parag Dhonde, M.E. (Computer), Datta Meghe College of Engineering, Navi Mumbai – 400708, India.
Parag.dhonde@gmail.com
Second Author – Asst. Prof. C. M. Raut, Datta Meghe College of Engg., Navi Mumbai, 400708, India
Cmr.cm.dmce@gmail.com

Correspondence Author – Parag Dhonde,
parag.dhonde@gmail.com, 9920906128.