CONTENT BASED IMAGE RETRIEVAL IN JAVA USING K-MEANS CLUSTERING ALGORITHM

[1]Shradha Kamshetty, [2]Shivani Kasar, [3]Mrunalini Kadu, [4]Prof.J.A.Nanajkar
[123]Student,Smt. Kashibai Navale College Of Engineering Pune,India
[4]Professor,Smt.Kashibai Navale College Of Engineering Pune,India

Abstract: The ever-increasing demand of the multimedia technologies have marked the need of large datasets of the imaging data. Once there is the huge dataset available it becomes equally important to retrieve the images according to our requirement and convenience. The traditional methods of retrieving the images using metadata has been bottleneck in the retrieval process. Also it is difficult to assign keywords to the huge amount of imaging datasets and remember likewise. A system is thus proposed for the effective and efficient retrieval of the desired images from the large databases using the closest match approach. The user can provide the image as a query input image. Visual characteristic of the image like shape is studied properly for the further image description and indexing purposes. The query image block is compared to all the image blocks in the databases one by one.
The images with certain similarity found are grouped together forming the clusters. Thus these clusters with some similar visual characteristics consists of the images that have the nearest distance with the input query image. The use of K-means clustering algorithm in many ways is efficient as the time elapsed by the system for indexing is reduced. The productivity of time as well as desired output results are obtained using K-means algorithm.

Keywords **– dataset, indexing, clusters, K-means.**

I. INTRODUCTION

Content-based image retrieval (CBIR)[2] also known as query by image content. It is the great deal of application in searching the images from large databases. CBIR became more favourable because web based image search engines were mostly based on metadata. A lot of garbage results as well as irrelevant images were obtained as outputs. If all the keywords were to be given manually it would just add the laborious task. It is also practically not feasible to capture every keyword that describes the image. Also, to get access to that keywords, its storage makes the system costly and inefficient. Thus system in which images based on their content can be filtered will accomplish both better indexing and also return more accurate results. "Content-based" as the word itself indicates means that the search will analyze the actual contents such as size, shape, angles, textures, colours, outlines, borders, etc of the image instead of the metadata such as keywords, tags ,or any kind of information associated with the image. 'Content' refers to the information that can be derived from the image itself. Data storage and image acquisition technologies have made many advancements which hand in hand has resulted in the creation of large image databases. To deal with these data, it is necessary to develop appropriate information systems to efficiently manage these collections. Image searching is one of the most important services that need to be supported by such systems. Two different approaches have been applied to allow searching on image collections: one based on image textual metadata and another based on image content information. The first retrieval approach is based on attaching textual metadata to each image and use traditional database query techniques to retrieve them by keywords [1],[2]. However, these systems require a previous annotation of the database images, which is a very laborious task. Furthermore, the annotation process is usually inefficient because users, generally, do not make the annotation in a systematic way. In fact, different users tend to use different words to describe a same image characteristic. The lack of systematization in the annotation process decreases the performance of the keyword- based image search. Conventional information retrieval is based solely on text, and these approaches to textual information retrieval have been transplanted into image retrieval in a variety of ways, including the representation of an image as a vector of feature values. However, “a picture is worth a thousand words.” Image contents are much more versatile compared with text, and the amount of visual data is already enormous and still expanding very rapidly. To cope with these special characteristics of visual data, content- based image retrieval methods have been introduced. In our system we have proposed two specific pages viz admin login page and user login page. Admin is the user who can monitor and control complete system. General user is any user can use our system whose concern is with image classification method. Tools used in our system are NetBeans refers to both a platform framework for Java desktop applications, and an integrated development environment (IDE) for developing with Java, JavaScript. Also, Swing application is being used.

.

II. BLOCK DIAGRAM OF EXISTING SYSTEM

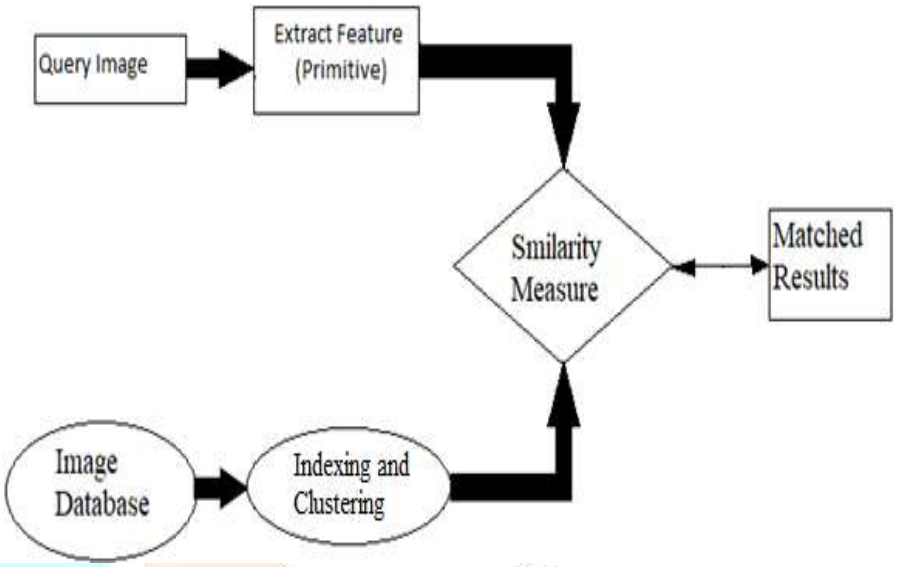


Fig 1:BLOCK DIAGRAM

III. RESEARCH METHODOLOGY

A. Problems with existing CBIR System

In CBIR systems with relevance feedback (RF), a user can mark returned images as positive or negative, which are again fed back into the systems as a query for the next stage of retrieval. The process repeats until the user is satisfied with the query result. Such systems are effective for many practical CBIR applications. There are two general types of image search: target search and category search. The goal of target search is to find a specific (target) image, such as a registered logo, a historical photograph, or a particular painting. The goal of category search is to retrieve a given semantic class or genre of images, such as scenery images or sky scrapers. In other words, a user uses target search to find a known image. In contrast, category search is used to find relevant images the user might not be aware ahead of time. An effective CBIR system, therefore, needs to have both an efficient search mechanism and accurate set of visual features. The Euclidean distances between the images reflect their similarity, and focus on finding new search techniques to improve the efficiency of target search. Existing target search techniques re- retrieve previously examined images (i.e., those retrieved in the previous iterations) when they again fall within the search range of the current iteration. This strategy leads to the certain disadvantages.

B. Proposed System

1. Feature Extraction: Feature extraction is one of the most important step in developing a classification system. This step describes the various features selected by us for classification of the selected image

2. Similarity Matching: The similarity matching stage is the main decision making stage of CBIR system and uses the features extracted in the previous stage to identify the image. It uses two algorithms Euclidean Distance Algorithm, K-Means Clustering Algorithm

3. Euclidean Distance Algorithm: It is necessary to have a certain measure to tell the quality of the system and to compare several images or words and establish which is the most similar image or word given a query. As it will be shown, distance methods play an important role in classification since some may be more suitable for certain features than others.

4. K-Means Clustering Algorithm: K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed a priori. The main idea is to define k centroids for k clusters, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done.

C. Proposed methodology

The results provided by search engines will already be quite good, especially since the queries chosen are most popular product queries for which large number of relevant web pages and images exists.

D. Features of the proposed system

The main feature of the existing system is the use of K-Means algorithm with does the Clustering Of Database is used which helps in proper indexing and saves much of the time. The K-means algorithm assigns each point to the cluster whose centroid (Centre) is in the nearest proximity (centroid is the average of all the points in the cluster). The centroid's coordinates are determined by calculating the arithmetic mean of all the points in the cluster separately for each dimension.

The biggest asset of this algorithm is its speed, which allow for it to be run on large databases. Also the simplicity of this algorithm makes it all the more easy to use.

There are two separate phases in the k-means algorithm- wherein the first phase involves the identification of k centroids, given that we know the number of clusters (K) beforehand thereby arriving at one centroid per cluster.

Initially K points which are likely to be in different clusters have to be selected, which are then made the centroids of their respective clusters.

The Euclidean distance is then calculated for each data point from each of the cluster centroid. Compare the values; find the closest centroid for the data point.

Then bind the centroid and the data point which results in the completion of the first phase and then an early grouping is done.

At this stage, the new centroids have to be determined by calating the mean value for each cluster.

After we get k new centroids, then a new binding is to be created between the same data points previously used and the nearest new centroid, thus generating a loop.

Because of this loop, the k centroids are prone to a change in their positions in a step by step manner.

This process is repeated until convergence criteria is meat means the centroids of clusters do not move anymore.

Steps in the K-Means Clustering Algorithm

1) Input:

$D = \{D_1, D_2, D_3, \ldots..,D_n\}$

D: set of n data items.

k: Number of desired clusters

2) Output:

A set of k clusters.

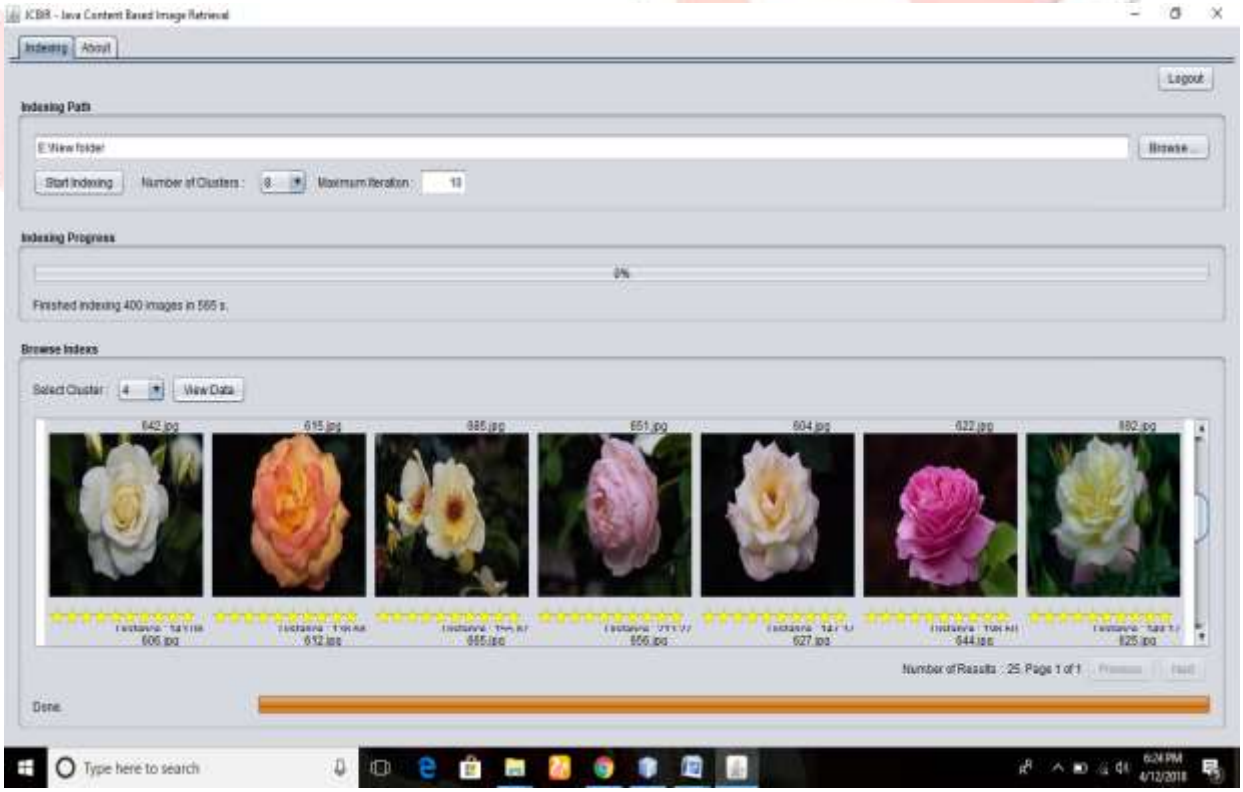
3) Steps:

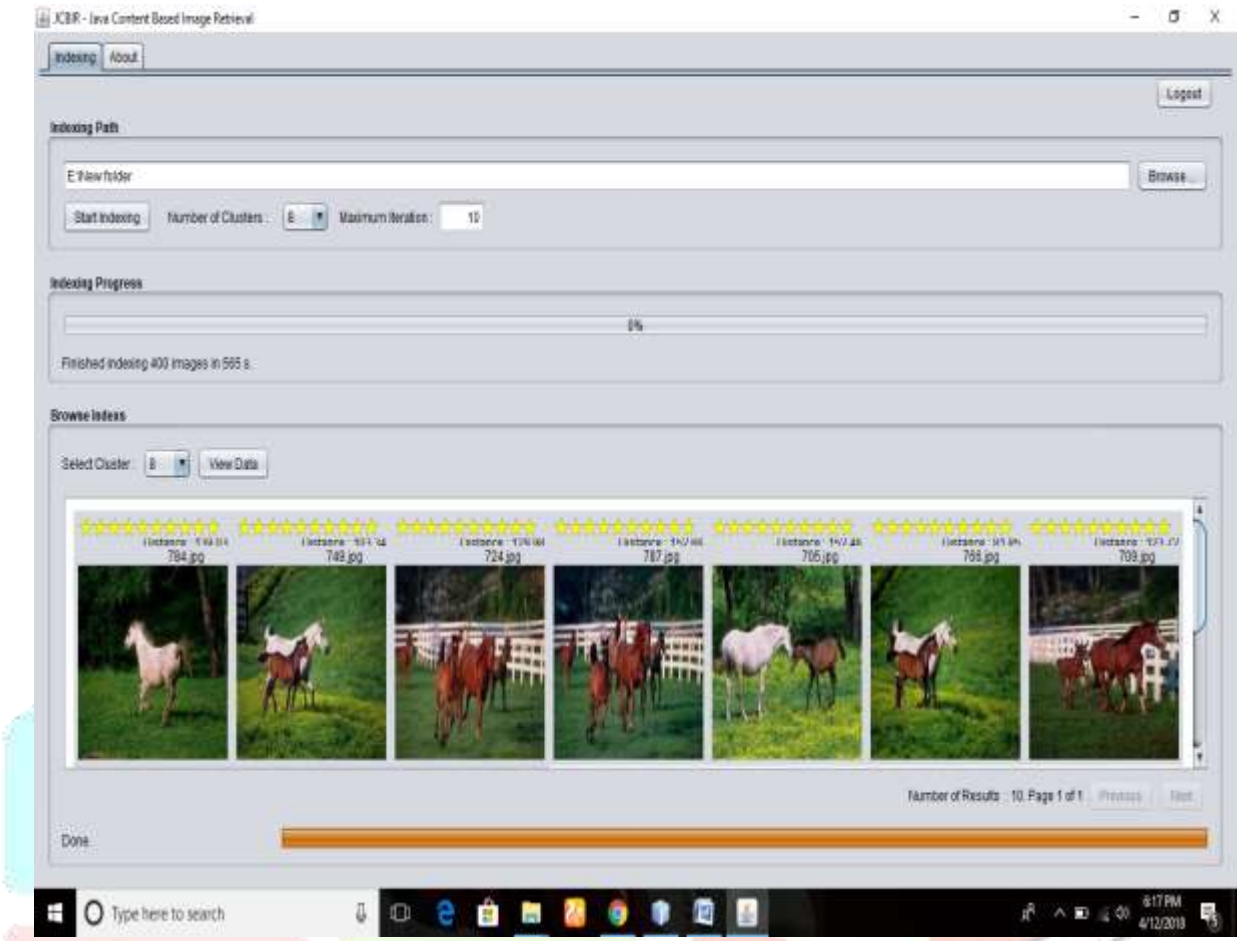
- Arbitrarily choose k data-items from D as initial centroids;
- Calculate the centroid by converting it into gradient format.
- Upload an image.
- Calculate the centroid of the image.
- Calculate the Euclidian distance of new image with the centroid of each existing cluster.
- Assign each data item D_i to the cluster which has the closest centroid.
- Calculate the new mean of each cluster.
- We have used the database of variety of images that include images of flowers, animals, buses, etc. The K-Means algorithm is used to group the images of similar characteristics together i.e form the clusters.
- Following are the visual description of how the cluster of similar instances are formed

1: The K-means Clustering algorithm collects the group of images with some common characteristics from the database, here shape being the feature used to form the cluster of flowers as shown below

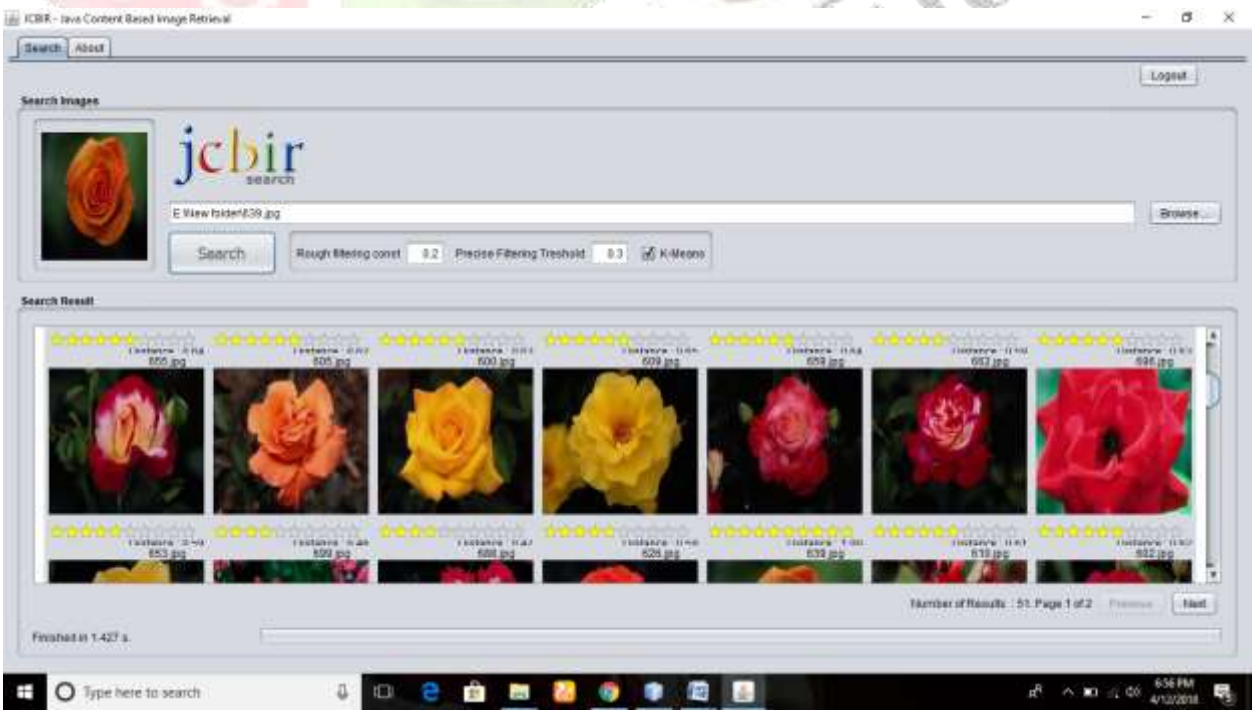
III. RESULTS

1. Clustering results of the system





2. **Final results of the Proposed System:**



IV. CONCLUSIONS:

1. While the problems of image retrieval in a general context have not yet been satisfactorily solved, the well-known artificial intelligence principle of exploiting natural constraints has been successfully adopted by system designers working within restricted domains where shape, color or texture features play an important part in retrieval.
2. Only in very specialist areas such as crime prevention has CBIR technology been adopted to any significant extent
3. This method of image retrieval removes the redundancy of conventional text based image retrieval by providing relevant search results and that too in a short time.

REFERENCES

[1] Datta R., Joshi D., Li J., and Wang J.Z., "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Comput Surv, vol. 40, no. 2, pp. 5:1-60, 2008.
[2] Alnihoud J., "Content-based Image Retrieval System Based on Self Organizing Map, Fuzzy Color Histogram and Subtractive Fuzzy Clustering," The International Arab Journal of Information Technology, vol. 9, no. 5, pp. 452458, 2012
[3] Hurtut T., Gousseau Y., and Schmitt F., "Adaptive Image Retrieval Based on the Spatial Organization of Colors," Comput Vis Image Und, vol. 112, pp. 101–113, 2008.
[4] Karthikeyan M. and Aruna P., "Probability Based Document Clustering and Image Clustering using Contentbased Image Retrieval," Appl Soft Comput, vol. 13, no. 2, pp. 959–966, 2013.
[5] Lin C. and Lin W., "Image Retrieval System Based on Adaptive Color Histogram and Texture Features," Comput J, vol. 54, no. 7, pp. 1136–1147, 2010.
[6] Lin C., Chan Y., Chen K., Huang D., and Chang Y., "Fast Color Spatial Feature Based Image Retrieval Methods," Expert Syst Appl, vol. 39, no. 9, pp. 11412–11420, 2011.
[7] Lin C., Chen R., and Chan Y., "A Smart Content-based Image Retrieval System Based on Color and Texture Feature," Image Vision Comput, vol. 27, no. 6, pp. 658– 665, 2009