Use of Genetic Algorithm Clustering Techniques for CBIR

Pooja Lather[1] Manjit Singh[2]
[1,2]Department of Computer Science and Engineering
[1,2]Ambala College of Engg. & Applied Research, Devsthali, Ambala, INDIA

Abstract— In this paper we first describe in brief content based image retrieval, feature extraction and then briefly survey the available literature on CBIR in which feature extraction clustering have been used. Brief introduction to the work that we have currently undertaken on use of CBIR is also given.

Key words: CBIR, GA

I. INTRODUCTION

It is Expected that there are 20 billion images in Imageshack and Facebook have 15 billion photos [13]. The third largest image warehouse on the Web seems to be News Corp's PhotoBucket, with 7.2 billion photos. after that Yahoo's Flickr tends to have 3.4 billion, including some videos.So due to Broad digitization of images over World Wide Web (www) the traditional keyword based search for image become inefficient for retrieval of images from huge dataset. CBIR Search Engine depends on the characterization of features i.e color, shape, and texture which can be automatically extracted from the images itself. CBIR approaches facilitate efficient and effective retrieval.

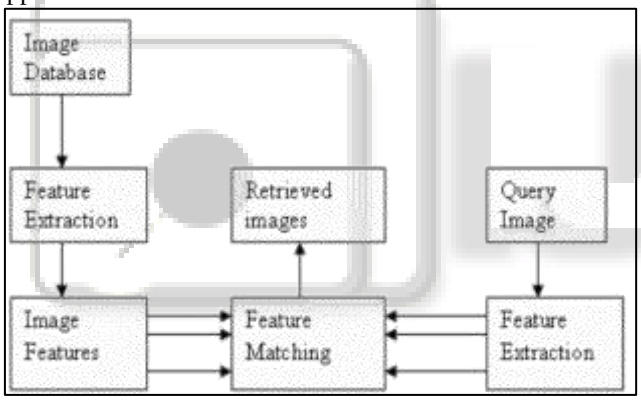

Fig. 1.1: Architecture of CBIR system

A. What is Feature Extraction?

Feature extraction is very vital step in image retrieval system to define the image with minimum number of descriptors. It is a means of extracting compact but semantically beneficial information from images which can define the image with its content. A "feature" means whatever that is localized, expressive and detectable.

1) Color Features

Color is the perception caused by the light as it interrelates with our eyes and brain. Humans tend to decide images based mostly on color features. Since of this, color features are the most extensively used in CBIR systems

2) Color space

To extract the color features from the content of an image, we have to select a color space and use its properties in the extraction. In common, colors are defined in 3-D color space.

3) Color Moments

Color moments are measures which can be used differentiate images built on their features of color. Once calculated, these moments offer a measurement for color similarity between images. These values of similarity can then be associated to the values of images indexed in a database for tasks like image retrieval. The basis of color moments sets in the assumption that the distribution of color in an image can be inferred as a probability distribution. Probability distributions are described by a number of exclusive moments. It therefore follows that if the color in an image trails a certain probability distribution, the moments can be used as features to categorize that image based on color.

B. Texture Features

In the arena of computer revelation and image processing, there is no clear-cut description of texture. This is because existing texture descriptions are based on texture study methods and the features extracted from the image. However, texture can be assumed of as recurring patterns of pixels over a spatial field, of which the addition of noise to the forms and their repetition frequencies consequences in textures that can seems to be random and unstructured. Texture assets are the visual patterns in an image that have assets of homogeneity that do not consequence from the occurrence of only a single color or intensity. The dissimilar texture properties as perceived by the human eye are, for example, consistency, directionality, smoothness, and coarseness, In real world acts, texture observation can be far more difficult. The numerous brightness intensities give increase to a blend of the different human perception of texture.

Image textures have convenient applications in image processing and computer vision. They include: recognition of image regions using texture assets, known as texture classification, recognition of texture boundaries using texture properties, known as texture segmentation, texture synthesis, and generation of texture images from known texture models. Since there is no recognized mathematical definition for texture, many different ways for computing texture features have been proposed over the years. Inappropriately, there is still no single method that works best with all types of textures.

C. Cluster-Based Retrieval Systems

It is more appropriate and beneficial to classify the images into clusters so to reduce the search domain in such search engines. Data Clustering is often taken as a step for speeding-up image retrieval and improving accuracy particularly in large databases.

D. Similarity

The similarity between two images is defined by a similarity measure. Selection of similarity metrics has a direct influence on the performance of content-based image retrieval. The kind of features vectors selected decides the kind of measurement that will be used to equate their similarity. If the features extracted from the images are defined as multi-dimensional points, the distances between

corresponding multi-dimensional points can be calculated. Euclidean distance is commonly used metric to measure the distance between two points in multi-dimensional space. Systems that use the Euclidean Distance are Netra, MARS, and Blobworld.

E. Performance Evaluation of the Retrieval Process

Evaluation of retrieval performance is a critical problem in content-based image retrieval (CBIR). Many different methods for calculating the performance of a system have been created and used by researchers. The most common calculation measures used in CBIR are precision and recall which are defined as,

$$Precision := Number\ of\ relevant\ images\ retrieved / Total\ numer\ of\ images\ retrieved$$

$$Recall := Number\ of\ relevant\ images\ retrieved / Total\ numer\ of\ images\ in\ database$$

F. A brief Review of Literature on use of CBIR, Feature Extraction, Clustering and GA.

Pengyu Hong. et.al., [1] had described an approach of classifying both positive and negative images for classification by applying Support Vector Machines (SVM). The SVM learning results are used to modify the preference weights for the relevant images. Their results shows that utilizing both positive and negative feedbacks can also lead to improved results.

Ujjwal Maulik et.al., [2] introduced GA based clustering. Clusters are optimized and searching capability of GA is exploited to search for clusters center based upon similarity metric.

D. Zhang et.al. [3] Have introduced a technique relating both color and texture features to increase retrieval performance. The color and texture features from the images, the database images are computed and indexed together. In the retrieval process, query image is given; images in the database are firstly ordered using color features. Then, in a second stage, top ordered images are selected and re-ranked as per their texture features. Two substitutions are provided to the user, first is the retrieval based on color features, and the second is retrieval based on combined features. If the first retrieval approach based methodology on color flops, the user will use the other alternative methodology i.e. the combined method of color and texture. Texture features degrades the system performance since they don't give the accurate description because these features are extracted globally from image.

P.Hiremath, et.al., [4] have introduced color, texture and shape features for CBIR. Color moments and moments on Gabor filter responses as local descriptors of color and texture respectively. While Shape information is captured in terms of edge images computed using Gradient Vector Flow field. Invariant moments are then used to record the shape features.

Anelia Grigorova et.al., [5] introduced adaptive retrieval approach for relevance feedback which launches a link between high-level concepts and low-level features, using the user's feedback not only to allocate proper weights to the features. The goal is to find a set of relevant features according to a user query whereas at the same time keeping a small sized feature vector to achieve better matching and lower complexity. The image description is updated during each retrieval by eliminating the least significant features and improved specifying the most significant ones. Results achieved on different image databases and two completely different feature sets show that the proposed algorithm outperforms previously proposed methods

Ricardo da S. Torres et.al., [6] introduced effectiveness of CBIR by combining features of images and weights are assigns to image similarities computed from the fusion of multiple features. They presented a genetic Programming framework for the combined similarity function in tis images are retrieved on the shape of object which is nonlinear combination of image similarities.

Ahmed Hosny El-Kholy et.al., [7] have introduced search engine known as SoM which is based on combined features and weighted similarity. The results obtained from SoM are efficient and respond quickly. Combining of more features in the SoM system is done to increase the efficiency of the results.

Yungang Zhang et.al., [8] introduced image retrieval approach depends on the extracted color and shape features. Vector Quantization (VQ) can offer a way of better exploiting the spatial information to produce different color histograms than scalar color quantization, thus VQ is employed in their work to extract color pattern of images. The shape feature of images is mined by curvelet transform, as it has been verified that the curvelet transform is an nearly optimal sparse representation of objects with edges. Combination of color and shape features is used and weighted by using Genetic Algorithm (GA), then used for image similarity measurement. Their Experimental results shows that the GA combined features can bring about good retrieval precision and speed simultaneously.

R.Balakrishnan et.al., [9] have introduced a Genetic Algorithm for clustering on image data. This technique normally has a long running time in terms of input set size. They proposed an efficient genetic algorithm for clustering on very large data sets, especially on image data sets.. Efficient time methods are used as a performance measure for clustering on image data. This paper compares Genetic algorithm with K-Means algorithm for clustering on image data.

Juli Rejito et.al., [10] have introduced K-Means clustering algorithm to develop clusters from each image database records, which later be used for optimizing image searching access period. The kept images in image database records are only restricted for the JPEG-type images. In this algorithm, cluster creation is based on maximum and minimum PSRN's (Peak Signal to Noise Ratio) calculation values from distinct records on a basic image and it will be treated as key image in each search for records with such cluster utilization.

Shrikant. et.al., [11] have introduced a way of using different feature descriptors such as, color ,texture and shape descriptors to signify low level features of image. There are the techniques called atrous wavelet transform (AWT) and Julesz's texton elements are used to produce the texton image. Also the multi texton histogram (MTH) is one of the techniques for these tasks. They integrate the advantages of co- occurrence matrix and histogram by signifying the attributes of co-occurrence matrix using histogram. User directed mechanism for CBIR using an interactive genetic algorithm (IGA) is proposed and implemented. The color characteristics such as the mean

value, standard deviation and image bitmap of a color image are used as features for retrieval.

Arvind Nagathan et al., [12] introduced a cbir system which make use of feed-forward back propagation neural network.at first step neural network is trained about the features of images. The image features reflected here are color histogram as color descriptor, GLCM (gray level co-occurrence matrix) as texture descriptor and edge histogram as edge descriptor. The training is done using back propagation algorithm. This trained when presented with a query image retrieves and shows the images which are relevant and similar to query from the database.

II. PROPOSED SYSTEM

A review of given literature shows that image search is usually done considering only one feature which can't provide good results.so it's important to consider multiple feature such as color ,texture .Searching the whole database will be time consuming therefore, images with similar features are grouped into related clusters. Images will be clustered using K-means and Genetic Algorithm .Clustering of images will be done off line before query processing so that to answer a query, the system doesn't require to search the entire image database. This method saves significant query processing time and computation load without compromising the retrieval precision in large database.

A. Methodology and Planning of Work

- Selecting Database: Database of N images i.e. Wang Database.
- Feature Extraction: We consider some important Multi-feature rather than single feature for higher accuracy.
- Building Index.
- Design of appropriate K-mean and Genetic algorithm and use this on image database after preprocessing.
- Comparison of performance measures by calculating precision and recall by k-means and GA.

1) Selecting Database

Database of N images i.e. Wang Database. These images contain 10 classes, each class contain 100 images.i.eAfricans,Beaches,Buildings,Buses,Dinosaurs,Elephants,flowers,Horses,Mountains,Foods

2) Feature Extraction

A feature Database using feature, Color histogram(32), Colorautocorrelogram(64), Color moments(6), Gabor wavelet(48), Wavelet moments(40).a database of feture set is prepared.

3) K-mean Clustering

K-means is applied to dataset we obtain after feature extraction so that similar images are grouped into one class. To speed up retrieval and similarity computation of our planned system, the database images are clustered using k-means clustering algorithm. We accomplish the clustering algorithm on the database as an offline step, and then we use these clusters to retrieve images relevant to the query image. This is complete by calculating the distance between the query image and the centroid of each cluster. The smallest distance between the query image and a centroid means that the query image is related to the centroid's cluster. Then, we calculate the distance between the query image and the images in that cluster to retrieve the most like images. To apply this, we chose some images arbitrarily from each class as queries.

4) Centroid Optimization using GAtool

The input data are characterized by coordinates $x_1, x_2, \ldots, x_K$ that characterize the objects. It is probable to define any number of clusters. The fitness function signifies the sum of squares of distances between the objects and centroids. The coordinates of centroids $c_{j1}, c_{j2}, \ldots, c_{jK}$ (j=1,2,…,M) are changed. The calculation allocates the objects to their centroids. The whole process is recurrent until the condition of optimum (minimum) of fitness function is reached. The process of optimization confirms that the defined coordinates $x_{i1}, x_{i2}, \ldots, x_{iK}$ (i=1,2,…,N) of objects and assigned coordinates $c_{j1}, c_{j2}, \ldots, c_{jK}$ of clusters have the minimum distances. The fitness function is stated by following formula (1):

$$f_{min} = \sum_{i=1}^{N} \min_{j=(1,2,\ldots,M)} \left(\sqrt{\sum_{i=1}^{K} (x_{ij} - c_{jl})^2}\right), \quad (1)$$

Where N is the number of objects, M the number of clusters and K dimension. This calculation is performed with help of GAtool command in MATLAB.

III. EXPERIMENTAL RESULTS

Comparison of precision and recall with k-means and optimized centroids using GA.

CLASS	Precision_km	Precision_GA
Africans	0.8409	0.7872
Beaches	0.6735	0.6818
Buildings	0.7045	0.8378
Buses	0.9767	0.9778
Dinosaurs	0.8929	1.0000
Elephants	0.9318	0.9767
Flowers	1.0000	1.0000
Horses	1.0000	1.0000
Mountains	0.9697	0.9750
Food	1.0000	1.0000
AVERAGE	0.89900	0.92363

Table 1: Comparison precision with k-means and GA

CLASS	Recall_km	Recall_GA
Africans	0.6400	0.6800
Beaches	0.7143	0.7800
Buildings	0.7561	0.8293
Buses	0.9318	0.9524
Dinosaurs	1.0000	1.0000
Elephants	0.9565	0.9762
Flowers	1.0000	0.9800
Horses	0.9792	1.0000
Mountains	0.9429	1.0000
Food	1.0000	1.0000
AVERAGE	0.89208	0.91979

Table 2: Comparison recall with k-means and GA

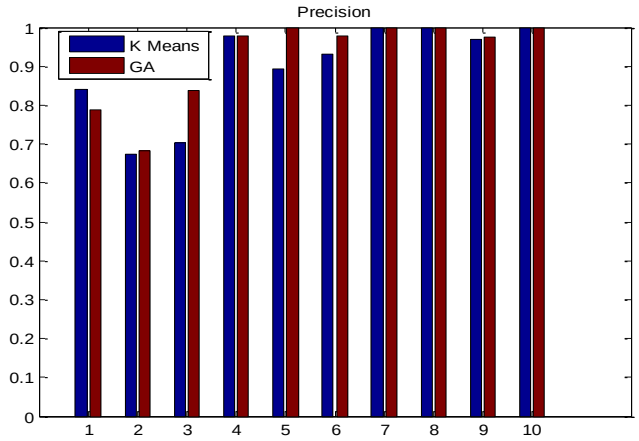

Fig. 2: Comparison precision with k-means and GA

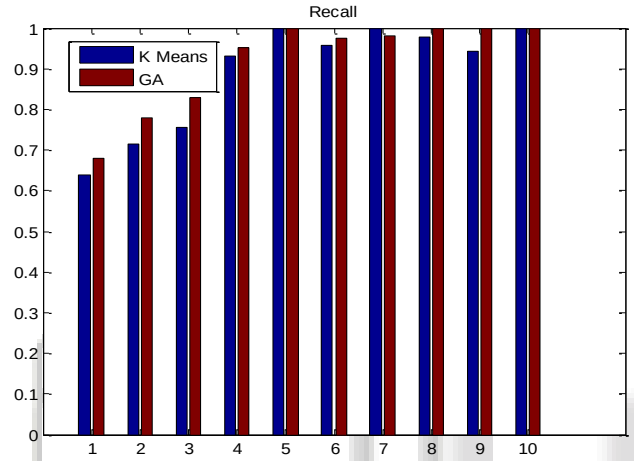

Fig. 3: Comparison recall with k-means and GA

IV. CONCLUSION

To the database first k-mean clustering is applied to find the closest cluster and relevant images. The centroids formed from the k-mean are than optimized using GAtool by defining fitness function. Now to the new optimized centroids again closest cluster is find and relevant images are searched in the closest cluster. The results shows that in recall there is increase in performance to 89% to 91% and in precision there is increase in 89% to 92%.Which shows are proposed system shows better results than previous system.

REFERENCES

[1] Pengyu Hong, Qi Tian, Thomas S. Huang "Incorporate support vector machines to content-based image retrieval with relevant feedback" .Appear in the proceedings of IEEE 2000 International Conference on Image Processing (ICIP'2000), pp. 750-753, Vol. 3, Vancouver, Canada, Sep 10-13, 2000

[2] Ujjwal Maulik Sanghamitra Bandyopadhyay," Genetic algorithm-based clustering technique". Elsevier 2000.

[3] D. Zhang [2004], "Improving Image Retrieval Performance by Using Both Color and Texture Features," Proceedings of IEEE 3rd International Conference on Image and Graphics (ICIG04), Hong Kong, China.

[4] P. Hiremath, and J. Pujari, "Content Based Image Retrieval using Color, Texture and Shape features" 15th International Conference on Advanced Computing and Communications, 2007.

[5] Anelia Grigorova, Francesco G. B. De Natale," Content-Based Image Retrieval by Feature Adaptation and Relevance Feedback".IEEE 2007.

[6] Ricardo da S. Torresa, Alexandre X. Falcãoa, Marcos A. Gonçalvesb, João P. Papaa, Baoping Zhangc, Weiguo Fanc, Edward A. Foxc, " A genetic programming framework for content-based image retrieval" 2008 Elsevier.

[7] Ahmed Hosny El-Kholy, Ahmed Mahmoud Abdel-Halei, Abdel-Rahman Hedar, "Content-Based Image Retrieval Using Combined Features and Weighted Similarity'' 2nd International Conference on Computer Technology and Development,2010.

[8] Yungang Zhang, Wei Gao, and Jun Liu, "Integrating Color Vector Quantization and Curvelet Transform for Image Retrieval" International journal of design , ANALYSIS and tools for circuits and systems, vol. 2, no. 2, august, 2011.

[9] Mr.R.Balakrishnan1, Mr.U.Karthick Kumar, "An Application of Genetic Algorithm with Iterative Chromosomes for Image Clustering Problems" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012.

[10] Juli Rejito, Retantyo Wardoyo, Sri Hartati, Agus Harjoko, "Optimization CBIR using K-Means Clustering for Image Database " International Journal of Computer Science and Information Technologies, Vol. 3 (4) , 2012.

[11] Shrikant Chavate , Prof. Vikas Gupta , " An approach used for user Oriented Content Based Image Retrieval using Interactive Genetic Algorithm" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013

[12] Arvind Nagathan, Manimozhi, Jitendranath Mungara "Content-Based Image Retrieval System using Feed-Forward Backpropagation Neural Network" IJCSNS International Journal of Computer Science and Network Security, VOL.14 No.6, June 2014.

[13] http://techcrunch.com/2009/04/07/who-has-the-most-photos-ofthemall-hint-it-is-not-facebook

