A Frame Work for Reducing the Time for Image Retrieval with Genetic Algorithm

S.SELVAM[1] and S.THABASUKANNAN[2]
[1]Research Scholar and Assistant Professor, Dept. of Computer Applications, N.M.S.S.V.N. College, Madurai-19.
[2]Principal, Pannai College of Engg& Tech, Sivagangai – 630 561, Tamilnadu, India.
E-Mail: s.selvammscmphil@gmail.com, thabasukannan@gmail.com

Abstract: In recent years especially in the last decade, the rapid development in computers, storage media and digital image capturing devices enable to collect a large number of digital information and store them in computer readable formats. The large numbers of images has posed increasing challenges to computer systems to store and manage data effectively and efficiently. Although this area has been explored for decade sand many researches have been done to develop some algorithms that solve some of its problems, no technique has achieved the accuracy of human visual perception indistinguishing images.

Nowadays, virtually all domains of human life including commerce, government, academics, hospitals, crime prevention, surveillance, engineering and historical research use information as images, so the volume of digital data is increasing rapidly. These images and their data are categorized and stored on computers and the problem appears when retrieving these images from storage media. Thus CBIR from large resources has become an area of wide interest in recent years especially in the last decade.

To retrieve any image, we have to search for it among the database using some search engine. Then, this search engine will retrieve many of images related to the searched one. The main problem fo r t he user is the difficulty of locating his relevant image in this large and varied collection of resulted images. To solve this problem, text-based and content-based are the two techniques adopted for search and retrieval.

The main objective of this paper is to build more generalized CBIR system which increase the searching ability and provide more accurate results. To improve the retrieval accuracy the system has taken the feedback from the user automatically. Here we used WANG database to evaluate the performance of the new system by calculating the precision and recall metrics. We also compared the new system with other existing CBIR systems. The performance of the new architecture in terms of average precision, recall and retrieval time has been shown to perform good. From the experimental results, it is evident that the new system has beaten other existing systems in terms retrieval time.

Keywords: CBIR, Genetic Algorithm, HARP Algorithm, Precision, Recall.

1. INTRODUCTION

The enormous collection of digital images on personal computers, institutional computers and Internet necessitates the need to find a particular image or a collection of images of interest. This has motivated many researchers to find efficient, effective and accurate algorithms that are domain independent for representation, description and retrieval of images of interest. There have been many algorithms developed to represent, describe and retrieve images using their visual features such as shape, color and texture. The visual feature representation and description play an important role in image classification, recognition and retrieval. The content based image similarity measurement algorithms, if chosen correctly for a particular image representation technique, will definitely increase the efficiency and effectiveness retrieval of data of interest.

Nowadays national geographic imagery archive has a size of Petabytes (PB) and grows to Terabytes (TB). It triggers the demand of qualitative and quantitative image retrieval systems. An image retrieval system is a computer based system for browsing, searching and retrieving images from a large database of digital images. Searching and retrieving is not bit by bit comparison. It is not a matching process on the raw data.

The drawbacks of the TBIR initiate to do the research in the field ofCBIR. In CBIR also known as query by image content (QBIC), retrieval is based on the image contents. Many techniques have been developed for the most important CBIR systems, which is a system, in which retrieves visual-similar images from large image database based on automatically derived image features, which has been a very active area recently. In most of the existing CBIR systems, the image content is represented by their low-level features such as colour, texture and shape. Thedrawback of low-level features is losing much detail information of the images, in case of looking for images that contain the same object or same scene with different viewpoints. In recent years, the interest

point detectors and descriptors are employed in many CBIR systems to overcome the above drawback.

Similarity can be defined as the quantitative measurement that indicates the strength of relationship between two image objects. Dissimilarity is also a quantitative measurement that reflects the discrepancy between two image objects.

In a CBIR system, the retrieval of images has been done by similarity comparison between the query image and all candidate images in the database. To evaluate the similarity between two images, the simplest way is to calculate the distance between the feature vectors representing the two images. To find more similar or relative images, the heuristic approach based Genetic algorithm has been used in the CBIR system.

Image retrieval techniques are useful in many image-processing applications. Content-based image retrieval systems work with whole images and searching is based on comparison of the query.

General techniques for image retrieval are color, texture and shape. These techniques are applied to get an image from the image database. They are not concerned with the various resolutions of the images, size and spatial color distribution. The content and metadata based system gives images using an effective image retrieval technique.

The primary goal of this paper is to reduce the computation time and user interaction. The conventional CBIR systems also display the large amount of results at the end of the process this will drove the user to spend more time to analyze the output images. In the proposed system we compute texture feature and color feature for compute the similarity between query and database images. This integrated approach will reduce the output results to a certain levels based on the user threshold value. The secondary goal is to reduce semantic gap between high level concepts and low level features. Generally the CBIR compute the similarity between the query image and the database images. Hence there might be chances for unexpected results at the end the retrieval process. The novel HARP cluster the output images and select one representative image from each clusters. A third goal is to evaluate their performance with regard to speed and accuracy. These properties were chosen because they have the greatest impact on the implementation effort.

Instead of Relevance Feedback we can use any clustering algorithm that based on the features extracted from the images themselves, and allocates those images into the nearest cluster. The algorithm calculates sand allocates until there is little variation in the movement of feature points in each cluster. Clustering is the unsupervised classification of patterns into groups. Its main task is to assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters.

In this paper, Color, Texture and Shape features were extracted and combined to form feature vector of image. For color features, the moments of the color distribution were calculated from the images and used as color descriptor. For texture features, we used Gabor filter, which is a powerful texture extraction technique in describing the content of image. For shape features, edge histogram features that include five categories were used as shape descriptor. These three descriptors were combined and optimized using GA with H A R P clustering accuracy as a fitness function to select optimum We performed GA with HARP clustering on the database as an offline step, and the system does not need to search the entire database images; instead just a number of candidate images are required to be searched for image similarity.

II. PREVIOUS STUDY

There are various approaches are present for CBIR. Some of the important literature which covers the more important CBIR System is discussed below.

a) Chin-Chin Lai et.al. have proposed an *interactive genetic algorithm* (IGA) to reduce the gap between the retrieval results and the users' expectation .They have used color attributes like the mean value, standard deviation, and image bitmap.They have also used texture features like the entropy based on the gray level co-occurrence matrix and the edge histogram.
b) Zhang Xu-bo et.al.have published a paper on *improved K-means clustering and relevance feedback* to re-rank the search result in order to remedy the rank inversion problem in CBIR. Experimental results show that the reranking algorithm achieves a more rational ranking of retrieval results and it is superior to Reranking via partial Grouping method
c) Lijun Zhao et.al.have proposed a *multi-round relevance feedback* (RF) strategy based on both support vector

machine (SVM) and feature similarity to reduce the gap between query and retrieve result.

d) SharadhRamaswamy et.al.have published a paper on a fast *clustering-based indexing technique*. In this method relevant clusters are retrieved till the exact nearest neighbors are found. This enables efficient clustering with low preprocessing storage and computation costs.

e) Nhu-Van Nguyen et.al. have proposed *Clustering and Image Mining Technique* for fast Retrieval of Images. The main objective of the image mining is to remove the data loss and extracting the meaningful information to the human expected needs. The clustering-repeat gives good result when the number of examples of feedback is small.

f) HuaYuan et.al. have presented a new *statistical model-based image feature extraction* method in the wavelet domain and a novel Kullback divergence-based similarity measure. The Gaussian Mixture Models(GMM) and Generalized GMM are presented to help extract new image features.

From the literature survey it is concluded that a wide variety of CBIR algorithms have been proposed in different papers. The selection feature is one of the important aspects of Image Retrieval

System to better capture user's intention. It will display the images from database which are the more interest to the user.

III. ARCHITECTURE OF NEW CBIR SYSTEM

Training image input:

The learning phase tells about the training process which a huge amount sample images are input in the first step. The genetic algorithm is used to train the features with different weights. For optimizing the feature weights and for fitness function, HARP algorithm is used. The training part outputs the classifying result and stores it in the feature database. All these steps performed offline and each class will be indexing along with its associated class ID in the index files.

Feature Extraction (Image signature): There are various kinds of visual features to represent an image, such as color, texture, shape, and spatial relationship. Since one type of features can only represent part of the image properties, a lot of work done on the combination of these features. The feature of each image is very much smaller in size compared to the image data, so the feature database contains an abstraction of the images in the image database.

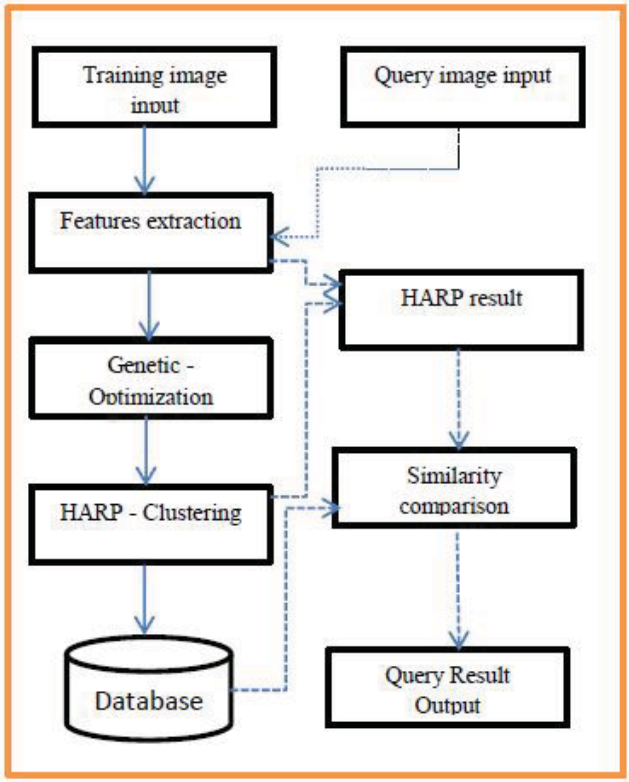

Fig. 1 : A New architecture for CBIR

a) **Colors:** are defined on a selected color space. Varieties of color spaces include, RGB, LAB, LUV, HSV (HSL), YCrCb and the HMMD. Common color features or descriptors in CBIR systems include color-covariance matrix, color histogram, color moments and color coherence vector storing, filtering and retrieving audiovisual data. The emerging MPEG-7 is a new multimedia standard, which has improved CBIR by providing a rich set of standardized descriptors and description schemas for describing multimedia content. MPEG-7 has included dominant color, color structure, scalable color, and color layout as color features. Here we used Color Structure Descriptor (CSD) as color feature.The CSD represents an image by both the color distribution of the image or image region and the local spatial structure of the color. CSD used a 8 × 8 structure to scan the total image. This descriptor counts the number of times a particular color is contained within the structuring element while the image or image region is scanned by this structuring element. It has used HMMD color space.

b) **Texture:** There exist different approaches to extract and represent textures. They can be classified into space-based, frequency-based models, and texture signatures. Some popular techniques i.e. wavelet transform, co-occurrence matrix, and Gabor filters are applied to express texture features for image.

c) **Shape:** It is seen that natural objects are primarily recognized by their shape. Two main types of shape feature are commonly used; *global features* such as aspect ratio, circularity and moment invariants and *local features* such as sets of consecutive boundary segments.

Genetic Algorithm–for optimization:

It is used to find approximate solutions to difficult-to-solve problems. It is search heuristic that mimics the process of natural selection. Here population of candidate solutions to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties which can be mutated and altered. It contains the following operators.

Selection: During each successive iteration, a proportion of the existing iteration is selected to breed a new generation. Individual solutions are selected through a *fitness-based* process, where fitter solutions are typically more likely to be selected. Certain selection methods rate the fitness of each solution and preferentially select the best solutions.

Mutation is a genetic operator used to maintain genetic diversity from one generation of a population of genetic algorithm chromosomes to the next.

Cross over is a process of taking more than one parent solutions and producing a child solution from them.

HARP –a Clustering algorithm: The algorithm forms clusters in a bottom-up manner. Initially put each article in its own cluster. Among all current clusters, pick the two clusters with the smallest distance. Replace these two clusters with a new cluster, formed by merging the two original ones. Repeat the above two steps until there is only one remaining cluster in the pool.In HARP algorithm, the accuracy level of clustering is more by using relevance indexing and merge score. The scalability level is also very high. The time taken for finding the closest cluster is very less.

Database: A database containing number of images with any one of the formats.bmp, .jpg, .tiff. is required.

Query: The user provides a sample image or sketched figure as the query for the system. This phase describes the images searching process. The user enters a query image for which the system extracts color, texture and shape features the features vectors of database images are previously extracted and stored.

Similarity Matching: Using the similarity metrics defined for color, texture and shape, the similarity distances between the query image and the centroid image of each class are calculated. The smallest distance (most similar) will determine to which the image belongs. The class with the smallest distance is returned and the images in this class will be compared with the query image.

Retrieval: The most matching images will be retrieved and then they are sorted in ascending order. The first *N* similar target images with smallest distance value to the query are retrieved and shown to the user.

IV. PERFORMANCEEVALUATION

Here we introduce the database that we select to test our system, and we also compare the new system results with some other existing CBIR systems. The images database that we used in our evaluation is WANG database. It is a subset of the Corel database of 1,000 images in JPEG format. 1,000 image databases went through our implemented system to extract the features and stored them. The extracted features are weighted by GA and they are used for classification by using the H A R P algorithm. The level of retrieval accuracy is a fa ct o r t o inf lu e nc e t he performance. In CBIR, the most commonly used performance measures are *Precision* and *Recall*. **Precision** is defined as the ratio of the number of retrieved relevant images to the total number of retrieved images. This means that precision measures the accuracy of the retrieval.

Recall is defined as the ratio of the number of retrieved relevant images to the total number of relevant images in the database. The recall measures the robustness of the retrieval.

In CBIR, if the precision score is 1.0 then every image retrieved by a search is Relevant. If the recall score is 1.0 then all relevant images are retrieved by the search is robust. We evaluate then n e w system b y u s i n g two metrics viz: the *Retrieval Effectiveness* and the *Retrieval Efficiency.*

a. ***RetrievalEffectiveness:*** A retrieved image is considered a match iff it is in the same class as the query image. The system works well and it retrieves b e t t e r results over the randomly selected images as queries by using GA and H A R P a l g o r i t h m .

b. **Retrieval Efficiency:** By assigning different weights to each feature to improve the efficiency we have used GA with a H A R P a l g o r i t h m to select optimum weights of features to g e t t h e accuracy.

Here by using clustering pre-process of the database image via H ARP algorithm decreases the average query response time, the similarity search time for image matching and increases the efficiency of the system.

Comparison of the new system with other existing systems

For each class in the database, we randomly selected 20 images as queries. Since we have 3classes in the database, we have 60 query images. For each query, we calculate the precision and recall of the retrieval. The average precisions and the average recall for each class based on the returned top 20 images were recorded. Moreover the new system result is compared against the performance of three methods.

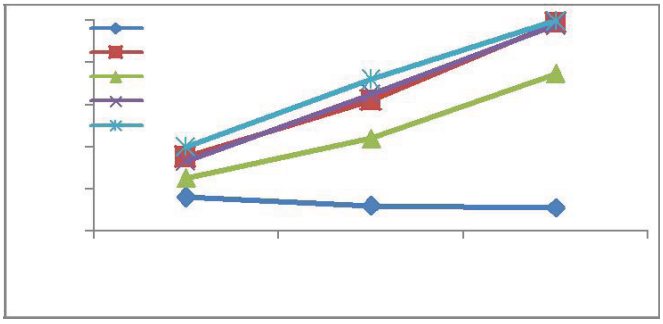

Figure 2: Comparison of Precision of the new system with various types of features.

The above figure shows that the new system performs significantly better in all three classes except elephant class.

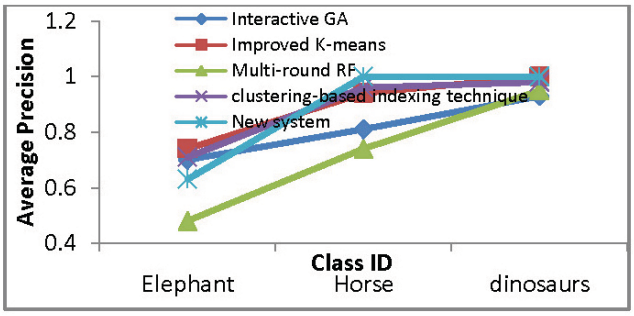

Figure3: Comparison of Precision of the n e w s y s t e m with some e xisting Systems

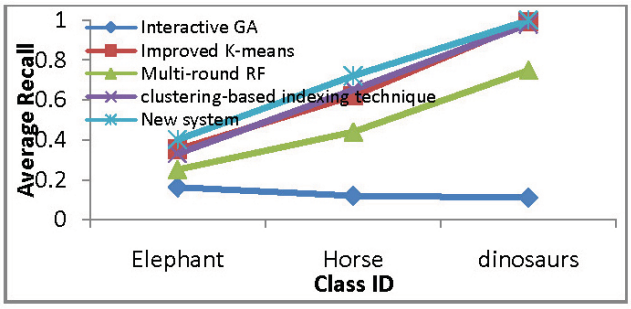

Figure4: Comparison of Recall of the new System with some existing Systems

The above figure shows that the new system performs significantly better than other systems, for all classes. This means that the new system can retrieve most of database images that match query image. The new system works well in the classification part of using GA with HARP algorithm. The average precision increased from 78.1%to88.2%, the average recall increased from 50.4%to69.9% and we obtained an average reduction in time equals 6.21seconds.

V. CONCLUSION

The explosive growth of image data leads to theneed of research and development of Image Retrieval.CBIR is currently a keen area of research in the area of multimediadatabases. Various research works had beenundertaken in the past decade to design efficient image retrieval techniques from the image databases. More précised retrieval techniques are needed to access the large image archives being generated, for finding relatively similar images. In this work the GA is combined with HARP clustering algorithm to improve the retrieval accuracy of the system. CBIR its aim is to retrieve semantically requested relevant and accur at e image concepts from large-scale image databases with lower computational time. In future enhancements we extend our features selections and introduce other distance measures to the user in order to improve the results.

Furthermore, we have improved the efficiency of the new system by not considering the whole database images for similarity computation but a number of candidate images were only considered. A candidate image is any database image that lies in the same cluster with the query image. The clustering process of the database images is performed by using H A R P c l u s t e r i n g algorithm. The results clearly proved the benefit of this clustering process in decreasing the retrieval time without sacrificing the retrieval accuracy.

REFERENCES

[1] V.GudivadaandV.Raghavan,"Content-basedImageRetrievalSystem"IEEEComputer, vol. 28, no 9, pp18-22, Sep. 1995.

[2] F.Long,H.Zhang,H.Dagan,andD.Feng, "Fundamentals of Content Based Image Retrieval," Multimedia Signal Processing Book, Chapter1,Springer-Verlag,BerlinHeidelberg NewYork,2003.

[3] R . C h a n g , J . H o , S . L i n , C . F a n n a n d Y . W a n g , " A N o velContentBasedImageRetrievalSystemusing K-means with Feature Extraction, "International Conference on Systems and Informatics,2012.

[4] I.El-Naqa, Y.Yang,N.Galatsanos, R.Nishikawa and M.Wernick ,"A Similarity Learning Approach to Content-Based Image Retrieval: Application to Digital Mammography," IEEE Transactions on Medical Imaging, 2009.

[5] B.WANG,X.ZHANG,andN.LI,"Relevance Feedback Technique For Content-Based Image Retrieval Using Neural Network Learning "Proceedings of the Fifth International Conference on Machine Learning and Cybernetics,Dalian, 2006.

[6] R.Datta,J.Li,andJ.Wang,"Content-BasedImageRetrieval-ApproachesandTrends of the New Age,"ACMComputing Surveys, vol. 40, no. 2, pp. 1-60,April 2008.

[7] J.HanandM.Kambr,"DataMiningConceptsandTe chniques,"2ndEd.,M organKaufmannPublisher, 2006.

[8] S.Selvam and Dr.S.Thabasu Kannan, "Design of an Effective Method for Image Retrieval", published IJIRAE, International Journal of Innovative Research in Advanced Engineering, Volume-1, March 2014, pp.51-56.

[9] P.JeyanthiandV.JawaharSenthilKumar, "Image ClassificationbyK-meansClustering."Advances in Computational Sciences and Technology, 2010.10. Petteri Kerminen, Moncef Gabbouj, The Visual Goodness Evaluation of Colors Based Retrieval Processes.

[11] M. J. Swain and D. H. Ballard, Indexing via Color Histograms, ICCV'90, pp. 390- 393, 1990.

[12] Tat-Seng Chua, Wai-Chee Low, and Chun-Xin Chu, Relevance feedback techniques for color-based image retrieval, In Proceedings of Multi-Media Modeling'98, IEEE Computer Society, pp 24-31, 2011.

[13] M. J. Swain and D. H. Ballard, Color indexing, International Journal on Computer Vision, vol. 7, no. 1, pp. 11--32, 2011.

[14] V. E. Ogle and M. Stonebraker, Chabot: retrieval from a relational database of images, IEEE Computer, vol. 28, no. 9, pp. 40-8, Sept. 1995.

[15] Michael Ortega, Yong Rui, Kaushik Chakrabarti, Sharad Mehrotra and Thomas S. Huang, Supporting Similarity Queries. In Proceeding of the ACM International Multimedia Conference, pp. 403-413, 2013.