International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 12 Issue 2 –JANUARY 2015.

AN INTEGRATION OF GENETIC ALGORITHM AND PROJECTED CLUSTERING FOR OPTIMIZATION OF CONTENT BASED IMAGE RETRIEVAL SYSTEM

S.Selvam, Dr.S.Thabasukannan

M.Sc., M.Phil., Assistant Professor, Research and Development Center, Bharathiar University, Coimbatore-641046
M.Phil., MBA., M.Phil.,M.Tech., Ph.D, Principal, Pannai College of Engg&Tech, Sivagangai – 630 561, Tamilnadu, India

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 the minicomputer 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 decades and many researchers have been do ne to develop some algorithms that solve some of its problems, no technique has achieved the accuracy of human visual perception indistinguishing images. To fulfill the latest technological requirements some domains like commerce, academic, health care, police etc., use information in the form of images.By using digitization, the volume of digital data is increasing tremendously. During retrieval of images from computer, some problems crop up. To solve this problem we use CBIR. 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 search done. The main problem for the 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.To evaluate the performance of new system, we use WANG database. The metrics used for evaluation are precision, recall and retrieval time. The performance can be evaluated by comparing some existing systems in CBIR. The performance of new system in terms of the metricsproves togood.**

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

I. INTRODUCTION

To meet the interested requirements of our collection of imagesfor future research, we have to search various built-in image libraries.To do the above in an effective manner some domain independent algorithms are required. Many algorithms are needed to represent, describe and retrieve images in an optimum manner. For this some data mining approach and heuristic approach are needed.The similarity checking algorithm is surely increases the efficiency in terms of retrieval. 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 retrievalsystems. An image retrieval system is a computer based system for browsing, searching andretrieving images from a large database of digital images. Searching and retrieving is not bit bybit comparison. It is not a matching process on the raw data.

The drawbacks of the TBIR initiate to do the research in the field of CBIR. 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 largeimage 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-levelfeatures such as colour, texture and shape. The drawback 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 manyCBIR 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 thequery image and all candidate images in the database. To evaluate the similarity between twoimages, 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 main aim of this new system is to minimize the computation time and user interaction. In conventional CBIR, the time taken to analyze the output images is more because the output displays at the end of the process. But in our newly constructed system, the time taken for analysis is meager, because it passes through various processing stages based on the user's threshold values.The step of this study is to reduce the gap between high and low level features as CBIR calculates the similarity between user query and repositories image. It may lead to unwanted retrieval of images. By using HARP, it groups the output images and a representative image from every cluster. The subsequent process is performance evaluation on the basis of speed and accuracy, because it gives strong impact on the implementation.

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 and 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 e-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 Gab or 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 c l u s t e r i n g accuracy as a fitness function to select optimum weights of features. We performed GA with H A R P c l u s t e r i n g on the database ea san offline step, and the system doesnot 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-boet.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) SharadhRamaswamyet.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) Hua Yuan 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 Generalised 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

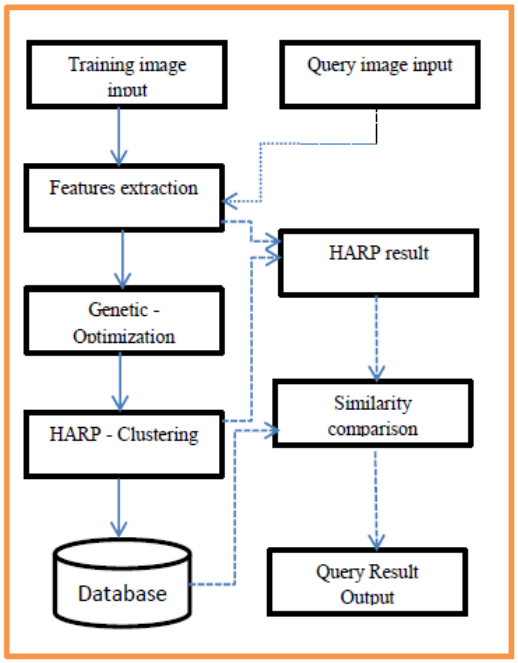

Figure 1 A New architecture for CBIR

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.

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 includecolor-covariance matrix, color histogram, color momentsand color coherence vectorstoring, 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 structuringelement 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 solution to complicated problems. It is based on heuristic approach that imitates the process of natural selection. It is used exclusively for the purpose of optimization. Each intermediary solution can be mutated and changed. It contains the following operators:

Selection: Duringeachiterationexisting input iterations are refreshed through a fitness process. If any iteration is best fit, then it is the solution and no further selection is needed.

Mutation is an interchange of data from one iteration to another.

Cross over is a process n-parent solutions used to derive a child solution.

HARP-a Clusteringalgorithm: The algorithm is based on bottom-up approach. Initially pick up each element among all current cluster on the basis of smallest distance by merging all the selected and related data on various clusters. 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. PERFORMANCE EVALUATION

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 database 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 HARP algorithm. The level of retrieval accuracy is a factor to influence the 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 the new system by using two metrics viz: the *Retrieval Effectiveness* and the *Retrieval Efficiency.*

a. ***Retrieval Effectiveness:*** A retrieved image is considered a match if it is in the same class as the query image. The system works well and it retrieves better results over the randomly selected images as queries by using GA and HARP algorithm.

b. **Retrieval Efficiency:** By assigning different weights to each feature to improve the efficiency we have used GA with a HARP algorithm to select optimum weights of features to get the accuracy.

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

Comparisonof the new systemwith otherexisting systems

For each class in the database, we randomly selected 20 images as queries. Since we have 3 classes 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.

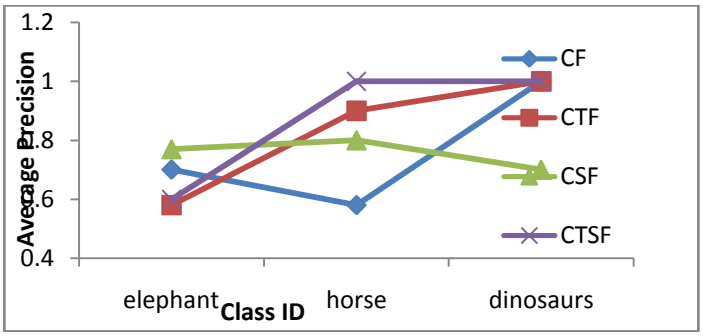

Figure 2: Comparison of Precision ofthenew system with various types of features.

The above figure shows that the new system performs significantly better in all three classes except elephant class. This result confirms that a fusion of multiple features can increase the performance of the system.

The below figure shows that the new system performs significantly better than other existing systems for all classes except elephant class. This is a good indicator for the effectiveness of our system. The reason behind the limitation in two classes is that those classes' images are very similar in term of the dominant color, texture and shape so, our new system may confused between them.

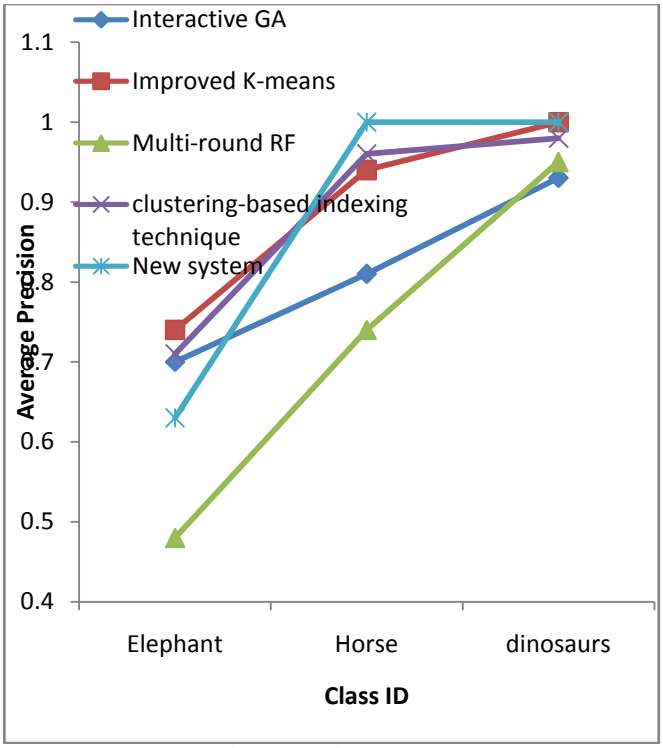

Figure3: Comparison of Precision of the new system with some existing Systems

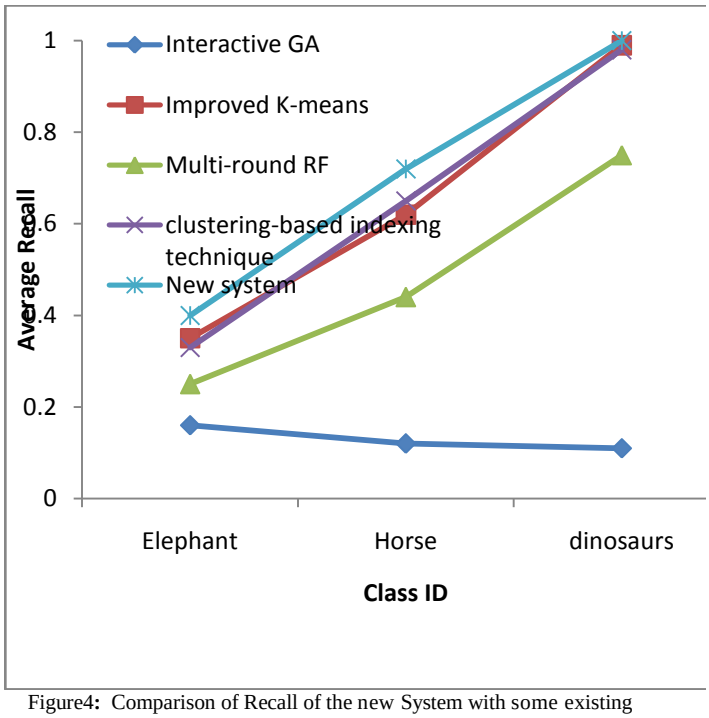

Figure4: Comparison of Recall of the new System with some existing s ystems

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 and the average recall increased from 78.1 % to 88.2 % and 50.4 % to 69.9 % respectively and obtained an average reduction in 6.21 seconds

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 imageretrieval techniques from the image databases. More précised retrieval techniques areneeded to access the large image archives beinggenerated, for finding relatively similar images. In thiswork the GA is combined with HARP clustering algorithm toimprove the retrieval accuracy of the system. Getting lower computational time and retrieving relevant and accurate image is possible by using CBIR. In future we have a proposal to disseminate the features selections and use other distance measures to improve the overall results. The efficiency of the new system is improved by considering candidate images for similarity computation purpose i.e. not considering the whole data base images. A candidate image lies in the same cluster with the query image the benefit of the clustering process clearly proved the retrievalaccuracy.

REFERENCES

[1] V.Gudivadaand V. Raghavan, "Content-based Image Retrieval Systems, "*IEEE Computer*, vol. 28, no 9, pp18-22, Sep. 1995.

[2] F.Long, H.Zhang, H.Dagan, and D. Feng,"Fundamentals of Content Based Image Retrieval," *Multimedia Signal Processing Book*, Chapter1, Springer-Verlag,Berlin Heidelberg New York, 2003.

[3] S.Selvam and Dr.S.ThabasuKannan, "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.

[4] S.Selvam and Dr.S.ThabasuKannan, "An Empirical Review on Image Retrieval System by using Relevance Feedback" proceeding of International Symposium on "Research innovation for quality improvement in Higher Education" conducted by Bharathiar University, Coimbatore, October 2014 and published in "Research and Trends in Data mining and Image Processing Technologies and Applications", Bloomsbury publishing India, London, New Delhi, New York, Sydney pp-1-11, October 2014, ISBN: 978-93-84052-11-9.

[5] R.Chang, J.Ho,S.Lin, C.Fannand Y.Wang, "A Novel Content Based Image Retrieval System using K-means with Feature Extraction, "*Internationa l Conference on Systems and Informatics*,2012.

[6] 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.

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

[8] R.Datta,J.Li,andJ.Wang,"Content-BasedImageRetrieval-ApproachesandTrends of theN ew Age,"*ACM Computing Surveys*,vol. 40, no. 2, pp. 1-60, April 2008.

[9] J.HanandM.Kambr,"DataMiningConceptsandTechniques,"2ndEd.,Morg anKaufmannPublisher, 2006.

[10] P. Jeyanthi and V.Jawahar Senthil Kumar, "Image Classification by K-means Clustering."*Advances in Computationa Sciences and Technology*, 2010.

1. First Author Personal Profile

Prof. S.SELVAM M.Sc., M.Phil, has been working as Assistant Professor and Head, Department of Computer Application, N.M.S.S.Vellaichamy Nadar College, Nagamalai, Madurai-19, Tamilnadu, India. He has more than 13 years of teaching experience. He has published three research papers in various refereed International/National Level Journals/Proceedings and Conference. His research paper also published in the book titled "Research and Trends in Data mining and Image Processing Technologies and Applications", Bloomsbury publishing India PVT, LD, London, New Delhi, New York, Sydney pp-1-11, October 2014, ISBN: 978-93-84052-11-9. Under his guidance two M.Phil scholar were awarded. His area of interest is Digital Image Processing.

2. Second Authour Personal Profile:

Prof.Dr.S.ThabasuKannan has been working as Professor and Principal in Pannai College of Engineering and Technology, Sivagangai and rendered his valuable services for more than two decades in various executive positions. He has published more than 50 research level papers in various refereed International/National level journals/proceedings. He has authored 11 text/reference books on the information technology. He has received 11 awards in appreciation of his excellence in the field of research/education. He has visited 5 countries to present his research papers/articles in various foreign universities. He has been acting as consultant for training activities at Meenakshi Trust, Madurai. His area of interest is Big data applications for bioinformatics applications. Under his guidance 8 Ph.d scholars pursuing and more than 150 M.Phil scholars were awarded. His several research papers have been cited in various citations.