IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-11 E-ISSN: 2347-2693

Parallel Indexing on Color and Texture Feature Extraction using R-Tree for Content Based Image Retrieval

L. Haldurai[1*] and V. Vinodhini[2]

[1*]*Department of Computer Science (PG), Kongunadu Arts and Science College, Coimbatore, Tamilnadu, India*
[2]*Department of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamilnadu, India*

www.ijcaonline.org

Received: Oct/26/2015 Revised: Nov /10/2015 Accepted: Nov/20/2015 Published: Nov/30/2015

***Abstract*—**Content Based Image Retrieval (CBIR) is a challenging method of capturing relevant image from a large storage space. This paper comprise of image features such as color and texture, which is intended to use in image retrieval. These features are extracted using fuzzy approaches. Numerous methods have been introduced in image retrieval systems. However, those methods have its drawbacks. In this paper novel system architecture for CBIR system which combines techniques includes CBIR and fuzzy based feature extraction, indexing procedure as well as genetic algorithm. This proposed approach is found to be very effective and efficient while comparing to previous methods and approaches in image retrieval in terms of retrieving most relevant images with less computational time.

Keywords*—Image Retrieval, Parallel indexing, Content Based image Retrieval (CBIR), R Tree, FCTH, Fitness Score*

I. INTRODUCTION

In recent years there is rapid development in technologies related to storage and digital image capturing devices. Due to increase in number of images challenges arise in the process of storing and managing data effectively and efficiently. Therefore the size of the database is also very difficult to handle and retrieve images. There has been a tremendous development in the process of developing tool and methods in retrieving the images from the large database.

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 encounters user here is the difficulty of locating his relevant image in this large and varied collection of resulted images. This problem referred to as *image retrieval problem*. To solve this problem, *text-based* and *content-based* are the two techniques adopted for search and retrieval in an image database [1].

II. CONTENT BASED IMAGE RETRIEVAL

Content-based image retrieval (CBIR), also known as query by image content (QBIC) is the application of computer vision techniques to image retrieval problem, that is, problem of searching for digital images in large databases [2]. It aims to finding images of interest from a large image database using the visual content of the images. "Content-based" means that the search will analyze the actual contents of the image rather than the metadata such as keywords, tags, and/or descriptions associated with the image. The term 'content' in this context might refer to colors, shapes, textures, or any other information that can be derived from the image itself [3].

In on-line image retrieval, the user can submit a query example to the retrieval system to search for desired images. The system represents this example with a feature vector and the distances (i.e., similarities) between the feature vectors of the query example and those of the image in the feature database are then computed and ranked. Retrieval is done by applying an indexing scheme to provide an efficient way of searching the image database. Finally, the system ranks the search results and then returns the results that are most similar to the query examples [4]. A typical Architecture for CBIR System is illustrated in Figure 1.

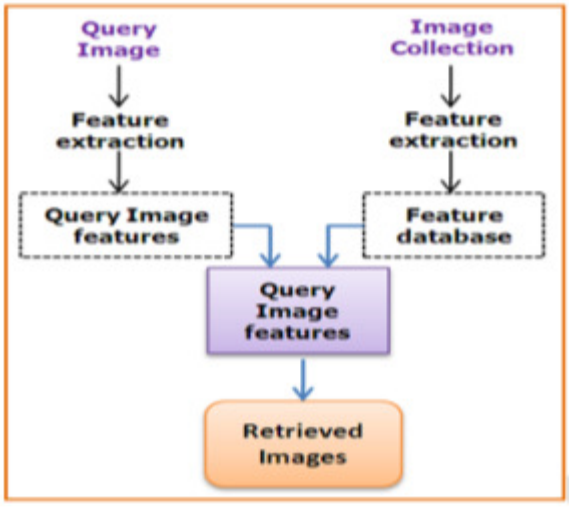

Figure 1: Typical Architecture of CBIR System.

Corresponding Author: *L. Haldurai,haldurai@gmail.com*

III. PROPOSED SYSTEM

In the proposed work an enhancement to basic content based image retrievaltechnique with indexing support by R-Tree technique. The enhanced feature helps inretrieving images from large database quickly. In this system an index is applied ondatabase of image features based on clustering technique. During this process k meansclustering concept uses features to find similarityamong the images. Based on similarity value the images are divided into clusters, thenthe new image which is to be verified with database is compared with these clustersand based on its similarity corresponding images in cluster are retrieved.The Experimental results show thatParallel Indexing on FCTH (Fuzzy Color and Texture Histogram) method is more efficient whencomparing with other methods. The proposed CBIR technique is evaluated by queryingdifferent images and the retrieval efficiency is evaluated by determining precisionvalues for the retrieval results.

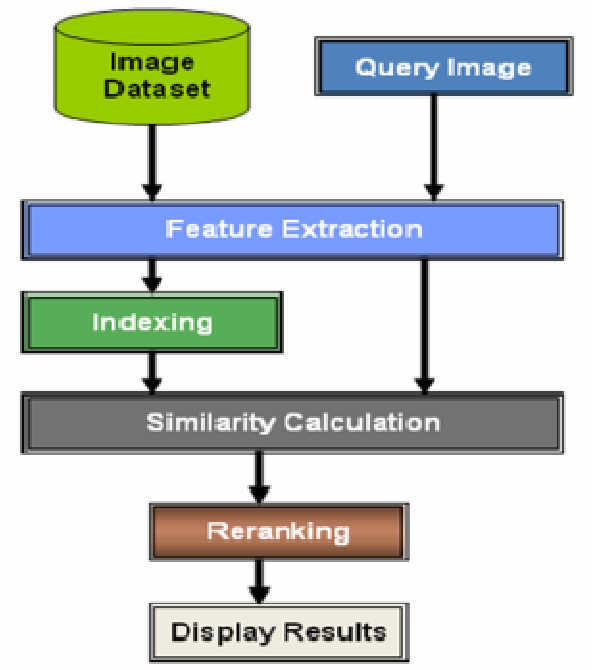

Figure 2: Proposed Architecture Diagram

IV. IMAGE DATASET INDEXING

Indexing is done using an implementation of the Document Builder interface. A simple approach is to use the Document Builder Factory, which creates Document Builder instances for all available features as well as popular combinations of features (e.g. all MPEG-7 features or all available features). A Document Builder is basically a wrapper for image features creating a Lucene Document from a Java Buffered Image. The signatures or vectors extracted by the feature implementations are wrapped in the documents as text. The document output by a Document Builder can be added to a Lucene index. Indexing methods that are in use on large image databases substantiated by performance figures: space partitioning, data partitioning, and distance-based techniques. In space-partitioning index techniques, the feature space is organized like a tree.

Data partitioning index techniques associate, with each point in feature space, a region that represents the neighborhood of that vector. An R-tree is such a data partitioning structure to index hyper rectangular regions in M-dimensional space. The leaf nodes of an R-tree represent the minimum bounding rectangles of sets of feature vectors. An internal node is a rectangle encompassing the rectangles of all its children. An R-tree is a variant which does not allow the minimum bounding rectangles in a node to overlap.

Feature based indexing techniques project an image as a feature vector in a feature space and index the space. The proposed feature based index structures R-tree. As mentioned above, R-trees are a direct extension of B-trees in k-dimensions. The data structure is a height-balanced tree which consists of intermediate and leaf nodes. Data objects are stored in leaf nodes and intermediate nodes are built by grouping rectangles at the lower level. Each intermediate node is associated with some rectangle which completely encloses all rectangles that correspond to lower level nodes. Considering the performance of R-tree searching, the concepts of coverage and overlap [8] are important. Coverage of a level of an R-tree is defined as the total area of all the rectangles associated with the nodes of that level. Overlap of a level of an R-tree is defined as the total area contained within two or more nodes. Obviously, efficient R-tree searching demands that both overlap and coverage be minimized.

V. FEATURE EXTRACTION

The extraction of a new low level feature that combines, in one histogram, color and texture information is named FCTH - Fuzzy Color and Texture Histogram. In FCTH [9] the input image is divided into 1600 blocks and histograms are formed using the fuzzy systems. The first fuzzy system is used for texture classification and the other fuzzy system is used for color classification. In Fuzzy Color Segmentation, the image is segmented in a preset number of blocks. Each block passes successively from all the fuzzy units. a set of fuzzy rules undertake the extraction of a Fuzzy Linking histogram [10]. This histogram stems from the HSV color space. Twenty rules are applied in a three-input fuzzy system in order to generate eventually a 10-bin histogram. Each bin corresponds to a preset color. Then proposes a two-input fuzzy system, in order to expand the 10-bins histogram into 24-bins histogram, importing thus information related to the hue of each color that is presented. In Fuzzy Texture Segmentation each image block is transformed with Haar Wavelet transform and a set of texture elements are exported. These elements are used as inputs in a third fuzzy

system which converts the 24-bins histogram in a 192-bins histogram, importing texture information in the proposed feature. With these values histograms are formed, now the texture information and color information are combined and the final output is quantized.

VI. RERANKING

The index iruns over the K-Example Images (EI) selected for re-ranking. Each of the chosen K images is used to re-rank the initial result list according to the similarity of the subset images to the K EIs. If K equals 10, the subset is re-ranked ten times, resulting in ten differently ranked lists, all containing the same images as the initial list. However, the ranking is based on the similarity between the low-level features of the K-EIs and the low-level features of all the images in the initially retrieved list.

Logically, for each list, the first image has to be the same as the EI with score 1.0 because it actually is the same image (EIs were taken from the initial list). As every image of the initial list, its maximum score found in any of the K score lists is finally assigned to the image. The set of images with a now new, re-assigned score is then sorted in decreasing order. As a consequence, the first K images used for re-ranking will also achieve the maximum score of 1.0. If the ten first images are taken, these ten images will still be the top-10, all having the same maximum score of 1.0. Only images on ranks higher than K may change.

Image clustering or categorization has often been treated as a preprocessing step to speed-up image retrieval in large databases and to improve the accuracy so that when a query is received, only a part of the database needs to be searched, while a large portion of the database may be eliminated in the search. Clustering is a way of grouping together data samples that are similar in some way according to some criteria that we pick its form of unsupervised learning. So, it is a method of data exploration – a way of looking for patterns or structure in the data that are of interest. They are presented with a set of data instances that must be grouped according to some notion of similarity. K-means clustering is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining the results. The algorithm converges when there is no further change in assignment of instances to clusters.

The Euclidean distance between points x and y, then the distance (d) from x to$y$is given by the formula:

$$d = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$

Re- Ranking process can be done on the features extracted from FCTH and in addition to that relevance information about the image is used. With this we can re-rank the similar images accordingly and retrieve the top matching images.

VII. IMAGE RETRIEVAL

For each query image the distance of it from each database image is calculated using a distance equation. If the distance is null then the images are exactly the same. As the distance increases it means that the similarity between the query image and the corresponding database image is less. Fitness function which is important component of genetic algorithm (GA) is used to find the best matching images for a query images.

Once we have extracted color and texture feature vectors from the query image, as well as the database images, we can use these feature vectors to measure the similarity between images in order to retrieve the most similar DB images to the query. The similarity between a query image q and a DB image d is defined by a distance between them, denoted as D(q,d), which is assessed according to the extracted color, texture and shape features. Two images are equivalent when thedistance value between them approaches zero and the similarity between them decreases as the distance value between them increases. This process helps us to rank the similar images accordingly and retrieve the top matching images.

A. Genetic Algorithm

The GAs is computer program that simulate the heredity and evolution of living organisms [11]. An optimum solution is possible even for multi modal objective functions utilizing GAs because they are multi-point search methods. Also, GAs is applicable to discrete search space problems. Thus, GA is not only very easy to use but also a very powerful optimization tool [12]. In GA, the search space consists of strings, each of which representing a candidate solution to the problem and are termed as chromosomes. The objective function value of each chromosome is called its fitness value. Population is a set of chromosomes along with their associated fitness. Generations are populations generated in an iteration of the GA [13].

The chromosomes that have best fitness value are retained while others are discarded. The chromosome that has the best fitness is chosen as the solution to the problem by repeating the process until one chromosome has best fitness value. Finally the most similar images relevant to the query image are retrieved from the image database.

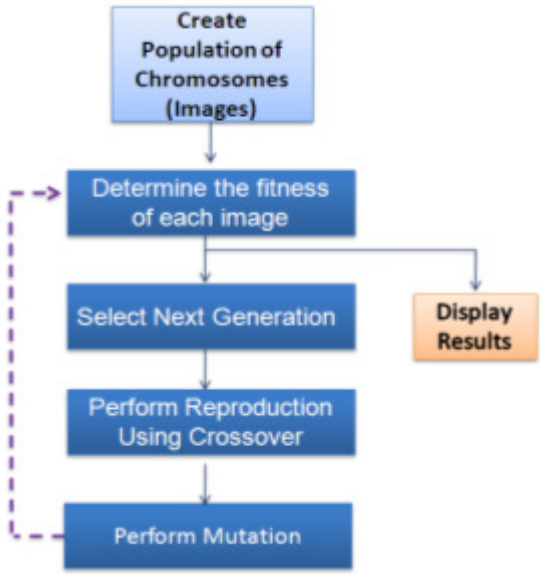

Figure 3: Genetic Algorithm based similarity measures

VIII. RESULTS AND DISCUSSIONS

The proposed CBIR system was tested using a Corel database of 1000 images that has 100 images in each category. The performance of each technique is measured by calculating its Image Retrieval Precision (IRP) and recall value as given in equation 1 and equation 2 respectively. For each query, the system collects database images which are similar to the query image. If the retrieved image belongs to same category as that of the query image, then we say that the system has appropriately identified the expected image, or else, the system has failed to find the expected image. Precision is the fraction of retrieved documents that are relevant to the search. Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system.

IRP = (No. of Relevant Image Retrieved / Total No. of Image Retrieved)

(Eq. 1)

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

Recall = (No. of Relevant Image Retrieved / No. of Relevant Images in Database)

(Eq. 2)

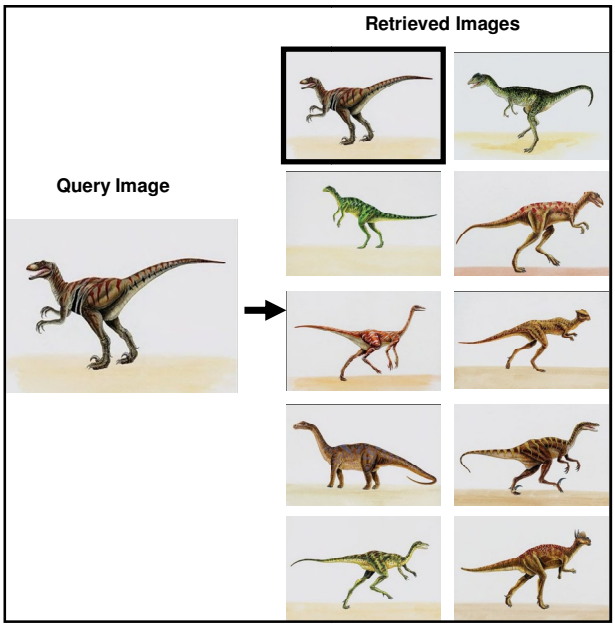

Figure 4: Top 10 retrieved images and the first image with black box matches with the query image.

Data set	Precision Value				
Query image category	Color Layout	Scalable Color	Edge Histogram	EMR	FCTH using R Tree
Natural scene	0.14	0.28	0.32	0.52	0.64
Butterfly	0.17	0.29	0.36	0.46	0.51
Car	0.11	0.25	0.33	0.53	0.68
Rose	0.13	0.29	0.38	0.58	0.71
Tree	0.15	0.23	0.35	0.65	0.84
Building	0.19	0.28	0.37	0.57	0.64
Elephant	0.14	0.24	0.31	0.71	0.97
Horse	0.18	0.25	0.39	0.59	0.78
Hills	0.13	0.26	0.34	0.64	0.88

Table 1: Comparison of precision value with different CBIR techniques

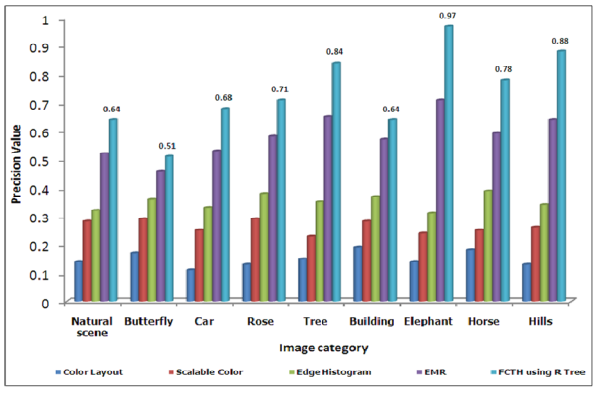

Figure 5: Performance analysis with different feature extraction techniques

To evaluate the performance of this system, precision value is used. The precision value is always in range 0 to 1 and the higher the value of this measure is the better the matching quality of the query image. The experimental result shows that the proposed feature improves the accuracy of image retrieval and computation time.

Data set	Computation Time				
Indexing	Color Layout	Scalable Color	Edge Histogram	EMR	FCTH using R Tree
Natural scene	3.5	3.2	3.1	2.7	1.8
Butterfly	3.8	3.6	3.1	2.4	1.5
Car	4.7	3.5	3.3	2.6	1.8
Rose	3.4	3.1	2.9	2.8	1.4
Tiles	4.6	4.2	3.3	2.4	2.1
Building	3.8	3.6	3.2	2.8	2.3
Tree	4.6	4.4	3.5	2.5	1.5
Horse	5.4	5.1	4.1	2.3	1.6
Hills	4.6	4.4	4.0	2.9	2.1

Table 2: Performance analysis Computation Time in (sec) with different feature extraction

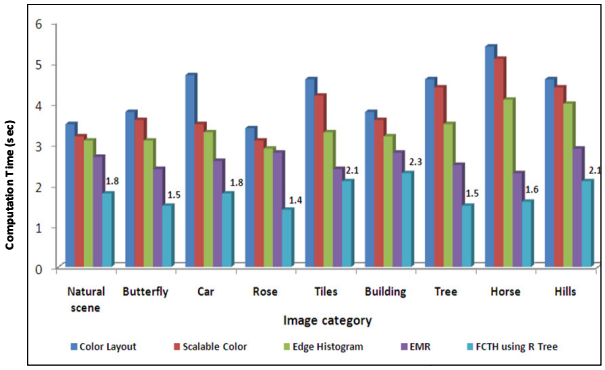

Figure 6: Computation Timeof different feature extraction Tecniques

IX. CONCLUSION

In this paper a novel approach of indexing the color and feature extraction of images and genetic algorithm has been implemented. The proposed method overcomes the shortcomings of previous methods. Experimental results reveal that parallel indexing can contribute in accurate image retrieval. Its main functionality is image-to-image matching and its intended use is for still-image retrieval. The evaluation criteria are provided by the GA and have been successfully employed as a measure to evaluate the efficacy of content-based image retrieval process. In future various other features can be considered to evaluate the efficiency of image retrieval.

REFERENCES

[1] S. Theodoridis and K. Koutroumbas, - Pattern Recognition,4th Edition, **2009**.

[2] M. Lew, N. Sebe, C. Djeraba and R. Jain, "Content-based Multimedia Information Retrieval: State of the Art and Challenges", ACM Transactions on Multimedia Computing, Communications, and Applications, Volume -02, Issue -01, Page No (1-19), February **2006**.

[3] 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, Volume -23, Issue -10, Page No (1233-1244), October **2004.**

[4] F. Long, H. Zhang, H. Dagan, and D. Feng, "Fundamentals of Content Based Image Retrieval, Multimedia Signal Processing Book, Chapter 1, Springer-Verlag, Berlin Heidelberg New York, Page No (1-26), **2003.**

[5] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds", Proc. Adv. NIPS, Volume- 16, Page No(169–176), **2003.**

[6] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang, "Mani foldranking based image retrieval", Proc. 12th Annu. ACM International Conference on Multimedia, Page No(9–16), **2004**.

[7] B. Xu et al., "Efficient manifold ranking for image retrieval",Proc. 34th Int. ACM SIGIR Conference on Research and Development in Information Retrieval, Page No(525–534), **2011**.

[8] M. Stonebraker, B. Rubenstein, and A. Guttman, ''Application of Abstract Data Types and Abstract Indices to CAD Data Bases,'' Tech. Report UCB/ERL M83/3, Electronics Research Laboratory, University of California, Berkeley, January **1983.**

[9] Savvas A. Chatzichristofis and Yiannis S. Boutalis "FCTH: Fuzzy Color and Texture Histogram – Low level feature for accurate image retrieval", IEEE DOI: 10.1109/WIAMIS(**2008**) Page No(191-196).

[10] S. Chatzichristofis and Y. Boutalis, "A Hybrid Scheme for fast and accurate image retrieval based on color descriptors", IASTED International Conference on Artificial Intelligence and Soft Computing (ASC 2007), Page No(280-285),August **2007.**

[11] Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", Addison-Wesley, **1989.**

[12] Hiroyasu T., "Diesel Engine Design using Multi-Objective Genetic Algorithm", Technical Report, Workshop on Design Environment, **2004.**

[13] Radwan A., Latef B., Ali A., and Sadek O., "Using Genetic Algorithm to Improve Information Retrieval Systems", World Academy of Science and Engineering Technology, Volume-17, Issue-2, Page No (6-13), **2006.**

AUTHORS PROFILE

L. Haldurai working as an Assistant Professor in Department of Computer Science (PG),Kongunadu Arts and Science College, Coimbatore. He received his Master of ComputerApplications (MCA) degree from Bharathiar University. His research interests are Data Mining,Data Communications and Artificial Intelligence & Expert Systems.

V. Vinodhinihas completed her Post Graduation and M.Phil degree from Bharathiar University. At present working as an Assistant Professor in Department of Information Technology, Dr.N.G.P. Arts and Science College, Coimbatore. Her research interests are Data Mining & Warehousing.