ISSN 2278-5612

IMAGE RETRIEVAL BASED ON WBCH AND CLUSTERING ALGORITHM

Asmita Shirsath[1], M. J. Chouhan[2], N. J Uke[3]
[1]M.E (I.T), SCOE, Pune,Pune University, India
shirsathasmita@gmail.com
[2]Assistant Professor, SCOE, Pune,Pune University, India
mrunal_rajput@yahoo.com
[3]Assistant Professor, SCOE, Pune,Pune University, India
nilesh.uke@gmail.com

ABSTRACT

Research on content-based image retrieval has gained tremendous momentum during the last decade. Color, texture and shape information have been the primitive image descriptors in content based image retrieval systems. In order to get faster retrieval result from large-scale image database ,we proposed image retrieval system in which image database is first pre-processed by Wavelet Based Color Histogram (WBCH) and K-means algorithm and then using Hierarchical clustering algorithm we index the previous result and then by using similarity measures we retrieve the images from pre-processed database. Experiments show that this proposed method offers substantial increase in retrieval speed but needs to be improved on retrieval results.

Indexing terms/Keywords

Color Histogram, DWT, Haar Discrete Wavelet Transforms,Hierarchical Clustering Algorithm, K-Means Algorithm.

Council for Innovative Research
Peer Review Research Publishing System
Journal: International Journal of Management & Information Technology

Vol. 5, No. 3
editor@cirworld.com
www.cirworld.com, member.cirworld.com

1. INTRODUCTION

World Wide Web (in short, WWW) has enabled users to access data in a variant of media formats. The number of digital images on the WWW is estimated to be more than hundreds of millions. This creates development of novel techniques for efficient storage and as well as for image retrieval. Content-based image retrieval (CBIR) aims at developing techniques that support effective searching and browsing of large image digital libraries on the basis of automatically derived image features[7].

At present, clustering has achieved great success in many fields including pattern recognition, system modeling, image processing, data mining and etc. Through clustering algorithm to classify large image database according to a similarity principle, similar images can be gathered together and thus the scope of image searching can be greatly reduced, and so the target image can be found quickly and accurately[1]. In order to get faster retrieval result from large-scale image database, we develop the image retrieval system based on an improved clustering algorithm for hierarchical indexing to image database based on the classical hierarchical clustering algorithm, K-means algorithm and the feature extraction method.

2. LITERATURE REVIEW

Research on content-based image retrieval has gained tremendous momentum during the last decade. A lot of research work has been carried out on Image Retrieval by many researchers, expanding in both depth and breadth. The term Content Based Image Retrieval (CBIR) seems to have originated with the work of Kato for the automatic retrieval of the images from a database, based on the color and shape present. Since then, the term has widely been used to describe the process of retrieving desired images from a large collection of database, on the basis of syntactical image features (color, texture and shape). The techniques, tools and algorithms that are used, originate from the fields, such as statistics, pattern recognition, signal processing, data mining and computer vision. CBIR is the most important and effective image retrieval method and widely studied in both academia and industry arena[2].

Content-based image retrieval, a technique which uses visual contents to search images from large scale image databases according to users' interests, has been an active and fast advancing research area since the 1990s. During the past decade, remarkable progress has been made in both theoretical research and system development. However, there remain many challenging research problems that continue to attract researchers from multiple disciplines[5].In the past decade, many image retrieval systems have been successfully developed, such as the IBM QBIC System, developed at the IBM Almaden Research Center, the VIRAGE System, developed by the Virage Incorporation, the Photobook System, developed by the MIT Media Lab, the VisualSeek System, developed at Columbia University, the WBIIS System developed at Stanford University, and the Blobworld System, developed at U.C. Berkeley and SIMPLIcity System[2].

The most closely related work, to our work, is of Cai-Yun Zhao[1]. They, however, consider ART2 algorithm for preprocessing the image database. Instead, in our work, we make use of feature extraction method i.e. use color and texture feature of images and k-means algorithm for preprocessing the image database. We also compare our work with [2], which describes the use of color and texture features of images. The key difference between is that the authors use color and texture features of images for image retrieval, instead we make use of this feature along with k-means algorithm for preprocessing the image database and then by using hierarchical clustering algorithm we form indexing of previous result, then by similarity measures we retrieve the similar images i.e., we build a hybrid system. Our system mainly consists of three steps: firstly we create initial cluster results by Pre-processing algorithm, next we establish hierarchical indexing based on the results of step one, then we retrieve the images by using Euclidean distance formula for similarity measures.

3. FEATURE EXTRACTION

The feature has the characteristics that describe the contents of an image.The feature is defined as a function of one or more measurements, each of which specifies some quantifiable property of an object, and is computed such that it quantifies some significant characteristics of the object[4]. Feature extraction is the process of generating features to be used in the selection and classification tasks. Typical low level features for content based image retrieval involves three major features which are color, shape and texture.In our system we make use of color and texture features for image retrieval.

3.1 COLOR

Color is the most extensively used visual content for image retrieval. Its three-dimensional values make its discrimination potentiality superior to the single dimensional gray values of images.Before selecting an appropriate color description, color space must be determined first[5].

3.2 COLOR SPACE

Each pixel of the image can be represented as a point in a 3D color space. Commonly used color space for image retrieval include RGB, Munsell, CIE L*a*b*, CIE L*u*v*, HSV (or HSL, HSB), and opponent color space. RGB images from the original images will be transformed to HSV color space. This color space is suitable since it reflect human perception and identification of color image. HSV stands for Hue, Saturation and Value. Hue defines the colors measured from 0-360, Saturation defines the concentration on color image (by percentage) and value defines the luminance of the image (by percentage).

3.3 COLOR HISTOGRAM

The color histogram serves as an effective representation of the color content of an image if the color pattern is unique compared with the rest of the data set. There are two types of color histograms, Global color histograms (**GCH**s) and Local color histograms (**LCH**s). A GCH represents one whole image with a single color histogram while the LCH divides an image into fixed blocks and takes the color histogram of each of those blocks. Figure 3.1 shows the GCH of an image[3].

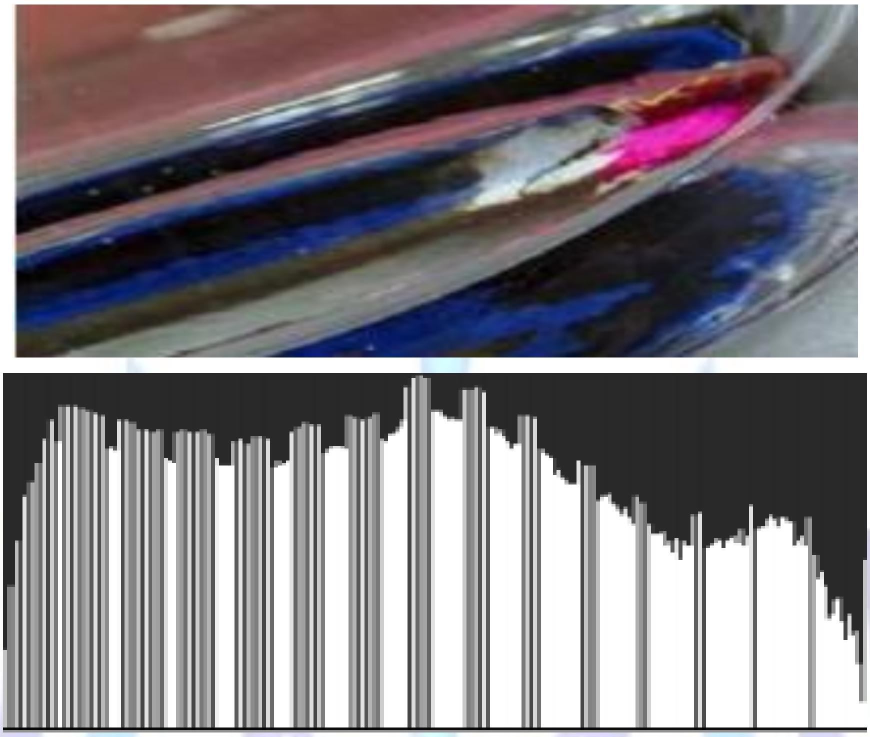


Figure 3.1: Image and its Global Color Histogram.

The color histogram is easy to compute and effective in characterizing both the global and local distribution of colors in an image. In addition, it is robust to translation and rotation about the view axis and changes only slowly with the scale, occlusion and viewing angle. We use Global Color Histogram in our proposed system.

3.4 TEXTURE

Texture is another important property of images. Various texture representations have been investigated in pattern recognition and computer vision .Figure 3.2 gives some examples of textured images[15].

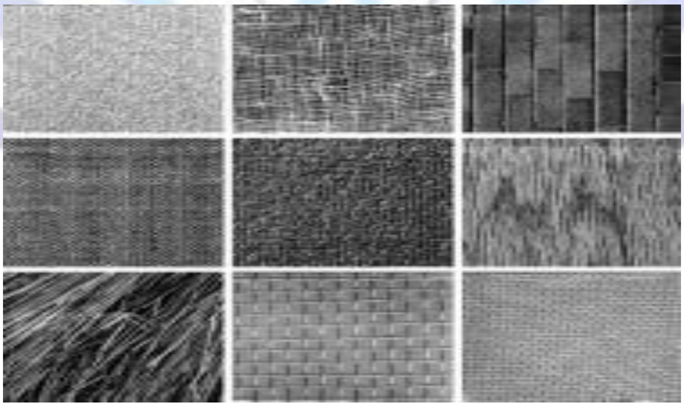


Figure 3.2: Images of oriented textures.

Basically, texture representation methods can be classified into two categories: structural and statistical. Structural methods, including morphological operator and adjacency graph, describe texture by identifying structural primitives and their placement rules. They tend to be most effective when applied to textures that are very regular. Statistical methods, including Fourier power spectra, co-occurrence matrices, shift-invariant principal component analysis (SPCA), Tamura feature, Wold decomposition, Markov random field, fractal model, and multi-resolution filtering techniques such as Gabor and wavelet transform, characterize texture by the statistical distribution of the image intensity [9, 10].

3.5 HAAR DISCRETE WAVELET TRANSFORMS

Discrete wavelet transformation (DWT) is used to transform an image from spatial domain into frequency domain. The wavelet transform represents a function as a superposition of a family of basic functions called wavelets. Wavelet transforms extract information from signal at different scales by passing the signal through low pass and high pass filters. Wavelets provide multiresolution capability and good energy compaction. Wavelets are robust with respect to color intensity shifts and can capture both texture and shape information efficiently. The wavelet transforms can be computed linearly with time and thus allowing for very fast algorithms. DWT decomposes a signal into a set of Basis Functions and Wavelet Functions. The wavelet transform computation of a two-dimensional image is also a multi-resolution approach, which applies recursive filtering and sub-sampling. At each level (scale), the image is decomposed into four frequency sub-bands, LL, LH, HL, and HH where L denotes low frequency and H denotes high frequency.

In our image retrieval system, we have used Haar wavelets to compute feature signatures, because they are the fastest to compute and also have been found to perform well in practice. Haar functions have been used from 1910 when they were introduced by the Hungarian mathematician Alfred Haar. The Haar transform is one of the earliest examples of what is known now as a compact, dyadic, orthonormal wavelet transform. The Haar function, being an odd rectangular pulse pair, is the simplest and oldest orthonormal wavelet with compact support [6]. Figure 3.3 shows a block diagram of Wavelet –Based Color Histogram i.e. feature extraction method used in our image retrieval system.

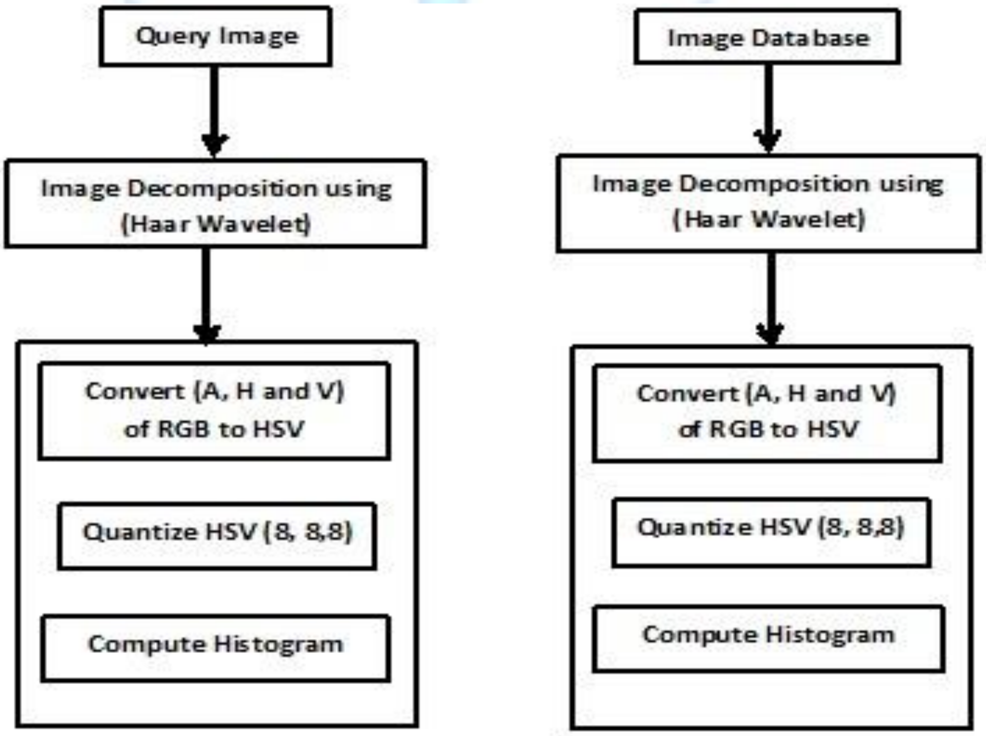

Query Image
Image Decomposition using (Haar Wavelet)
Convert (A, H and V) of RGB to HSV
Quantize HSV (8, 8,8)
Compute Histogram
Image Database
Image Decomposition using (Haar Wavelet)
Convert (A, H and V) of RGB to HSV
Quantize HSV (8, 8,8)
Compute Histogram

Figure 3.3: Block Diagram of Wavelet –Based Color Histogram (A-Approximate coefficient, Horizontal detail coefficient, V-vertical detail coefficient)

4. CLUSTERING METHODS

Clustering is an important analysis tool in many fields, such as pattern recognition, image classification, biological sciences, marketing, city-planning, document retrievals, etc. Clustering is the process of partitioning or combination a given set of patterns into displaces clusters. Most of the clustering techniques fall into two major categories, and these are the hierarchical clustering and the partitional clustering. Hierarchical clustering can further be divided into agglomerative and divisive, depending on the direction of building the hierarchy. Hierarchical techniques produce a nested sequence of partitions, with a single all inclusive cluster at the top and singleton clusters of individual objects at the bottom. K-means, which is one of the representative partitioning methods, obtains the number of clusters through minimizing the objective function. K-means has higher efficiency compared with the hierarchical methods. However, the number of clusters K needs to be fixed iteratively. Thus, K-means is often required to be run many times and is computationally expensive. How to determine the number of clusters becomes an increasingly important problem. The common trail-and-error method generally depends on certain clustering algorithms and is inefficient when the dataset is large. At present, several existing clustering algorithms focus on combination of the advantages of hierarchical and partitioning clustering algorithms. In our system we have also focus on combination of the advantages of hierarchical and partitioning clustering algorithms[11, 12, 13, and 14].

5. SIMILARITY MEASURES

Many similarity measures have been developed for image retrieval based on empirical estimates of the distribution of features in recent years. Performance of an image retrieval system is dependent on the type of similarity measures used[3]. Instead of exact matching, content-based image retrieval calculates visual similarities between a query image and images in a database. Accordingly, the retrieval result is not a single image but a list of images ranked by their similarities with the query image. Many similarity measures have been developed for image retrieval based on empirical estimates of the distribution of features in recent years. Different similarity/distance measures will affect retrieval performances of an image retrieval system significantly[5]. For clustering numeric field there are many well-known methods such as Euclidean distance, Minkowski distance, Manhattan (City-Block), etc., but all the distance measures discussed yields the same result for 1-norm distance. So, Euclidean Method is selected for this system. Euclidean Distance is the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. The Euclidean distance between points P= (p1, p2,… pn) and Q=(q1,q2,…qn), in Euclidean n-space, is calculated using:

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Where, p_i is the data point in x-axis, q_i is the data point in y-axis[11].

6. PROPOSED METHOD FOR IMAGE RETRIEVAL SYSTEM

6.1 IMAGE RETRIEVAL SYSTEM ARCHITECTURE

Image retrieval is generally conducted by searching the most similar images from databases to the query image. The overall image retrieval system is shown in Figure 6.1

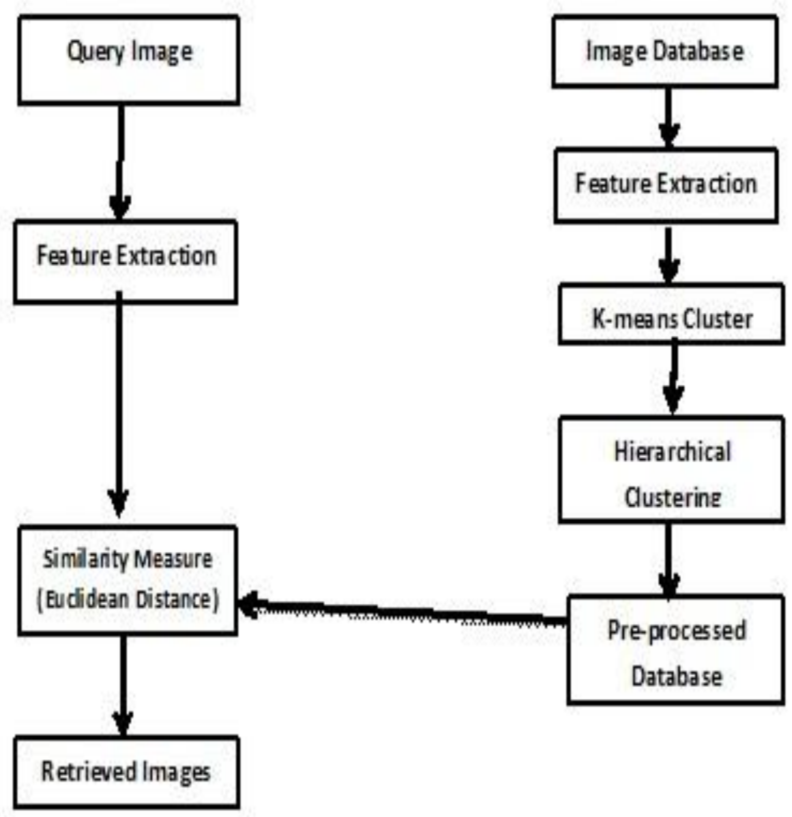


Figure 6.1: Overall Proposed Image Retrieval System

6.2 SYSTEM PROCEDURE

Step1: Extract the Red, Green, and Blue Components from an image.

Step2: Decompose each Red, Green, Blue Component using Haar Wavelet transformation at 1st level to get approximate coefficient and vertical, horizontal detail coefficients.

Step3: Combine approximate coefficient of Red, Green, and Blue Component.

Step4: Similarly combine the horizontal and vertical coefficients of Red, Green, and Blue Component.

Step5: Assign the weights 0.003 to approximate coefficients, 0.001 to horizontal and 0.001 to vertical coefficients (experimentally observed values).

Step6: Convert the approximate, horizontal and vertical coefficients into HSV plane.

Step7: Color quantization is carried out using color histogram by assigning 8 level each to hue, saturation and value to give a quantized HSV space with 8x8x8=512 histogram bins.

Step8: The normalized histogram is obtained by dividing with the total number of pixels.

Step9: Repeat step1 to step8 on an image in the database.

Step 10: Apply K-means algorithm to obtain group of cluster of feature vectors.

Step 11: Then hierarchical index will be firstly established by Hierarchical clustering algorithm on the results of previous step, and then retrieval will be done based on the indexing. We use the Euclidean distance formula as similarity measurement.

6.3 IMAGE RETRIEVAL METHOD

Image retrieval is generally conducted by searching the most similar images from databases to the query image. In our system, hierarchical index will be firstly established by the algorithm proposed for image database, and then retrieval will be done based on the indexing. Figure 6.2 shows the image retrieval flowchart.

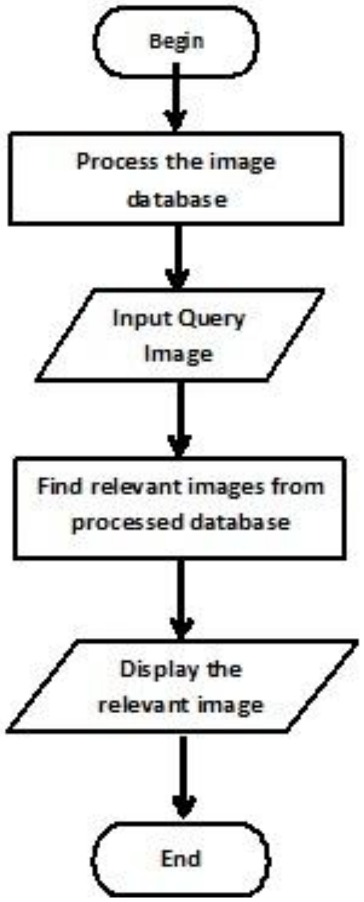

Begin
Process the image database
Input Query Image
Find relevant images from processed database
Display the relevant image
End

Figure 6.2: Image Retrieval Flowchart

6.4 USER SCENARIO

Use case diagrams are used to model the static use case view of a system. We can model context of a system or requirement of a system using use case diagram. Figure 6.3 shows the user scenario of the system.

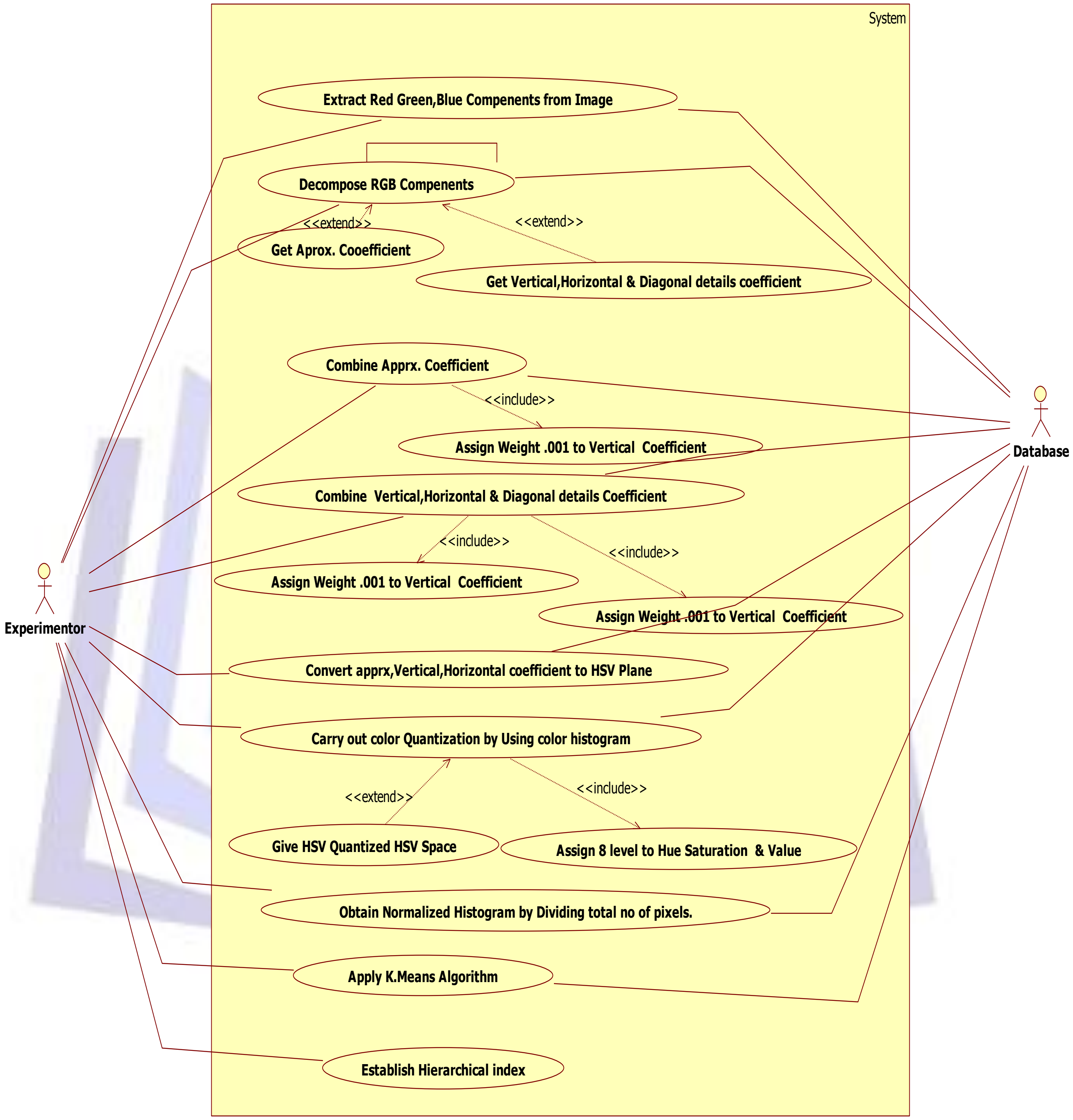
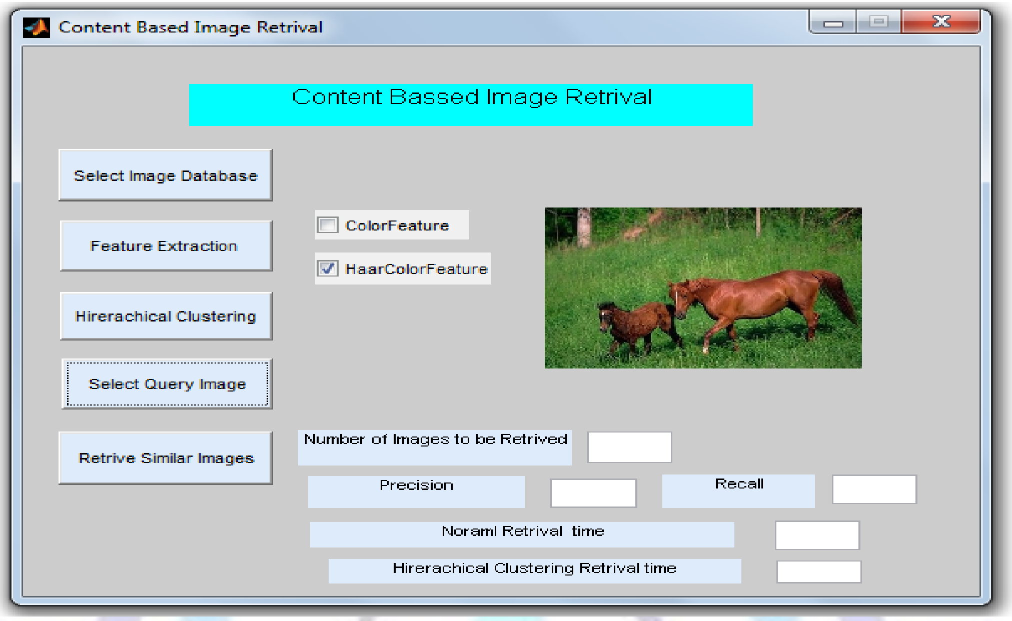

System
Extract Red Green,Blue Compenents from Image
Decompose RGB Compenents
<<extend>>
Get Aprox. Cooefficient
<<extend>>
Get Vertical,Horizontal & Diagonal details coefficient
Combine Apprx. Coefficient
<<include>>
Assign Weight .001 to Vertical Coefficient
Combine Vertical,Horizontal & Diagonal details Coefficient
<<include>>
<<include>>
Assign Weight .001 to Vertical Coefficient
Assign Weight .001 to Vertical Coefficient
Convert apprx,Vertical,Horizontal coefficient to HSV Plane
Carry out color Quantization by Using color histogram
<<extend>>
<<include>>
Give HSV Quantized HSV Space
Assign 8 level to Hue Saturation & Value
Obtain Normalized Histogram by Dividing total no of pixels.
Apply K.Means Algorithm
Establish Hierarchical index
Experimentor
Database

Figure 6.3:User Scenario for proposed Image Retrieval System.

6.5 USER INTERFACE DESIGN

User can use following user interface using which he can retrieve relevant images for the query image and see the performance of our image retrieval system.User is provided with options such as selecting image database, feature extraction, hierarchical clustering, uploading query image and then retrieving relevant images from database for the query image.Also we compare our system with other method i.e Color Histogram.We also calculate the Precision, Recall and

Normal retrieval time and the retrieval time required for our system. Figure.6.4 shows user interface for proposed Image Retrieval System



☐ ColorFeature

☑ HaarColorFeature

Figure.6.4 User interface for proposed Image Retrieval System.

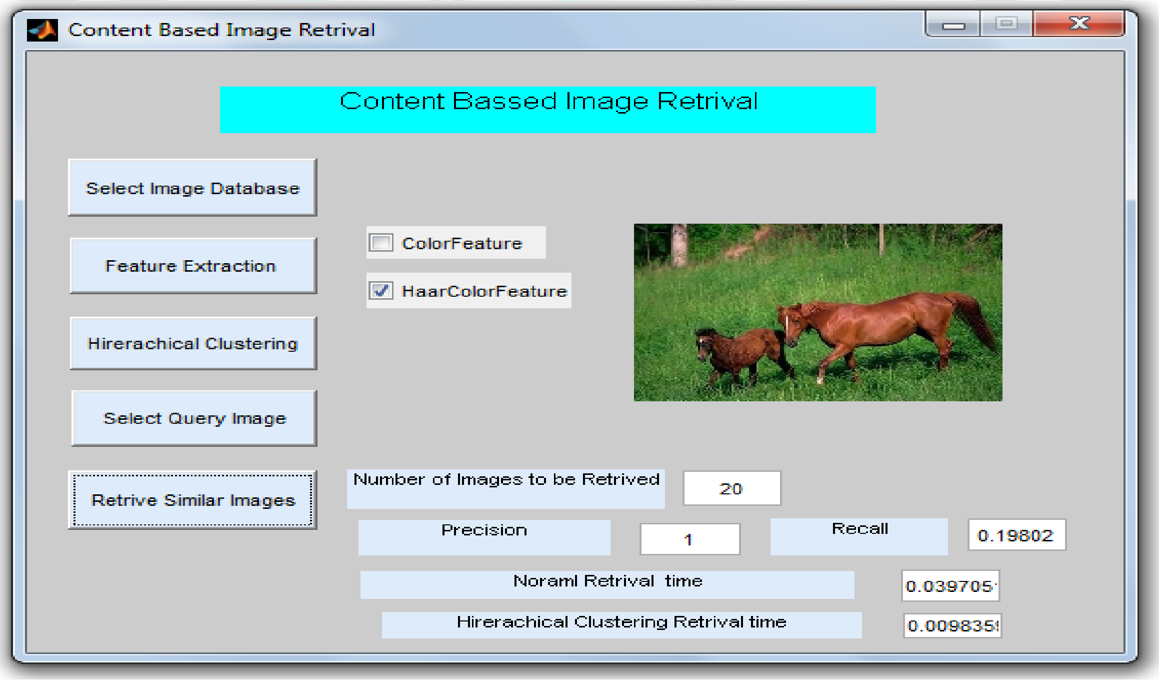
7. EXPERIMENT SETUP AND RESULTS

7.1 EXPERIMENT SETUP

The proposed image retrieval method has been implemented using Matlab 10. We used a general-purpose WANG database containing 1,000 images; in JPEG format of size 384x256 and 256x386 [8]. The search is usually based on similarity rather than the exact match.We retrieves the first 20 similar images.

7.2 EXPERIMENTAL RESULTS

The experiments for our proposed system is carried on query images i.e. on horse image and the results are as shown below. Figure 7.1 shows the user interface for query image of horse and results for this query image of horse is shown in Figure 7.2.



☐ ColorFeature

☑ HaarColorFeature

Number of Images to be Retrived 20

Precision 1

Recall 0.19802

Noraml Retrival time 0.039705

Hireachical Clustering Retrival time 0.009835

Figure 7.1: Query Image-horse for proposed Image Retrieval System.

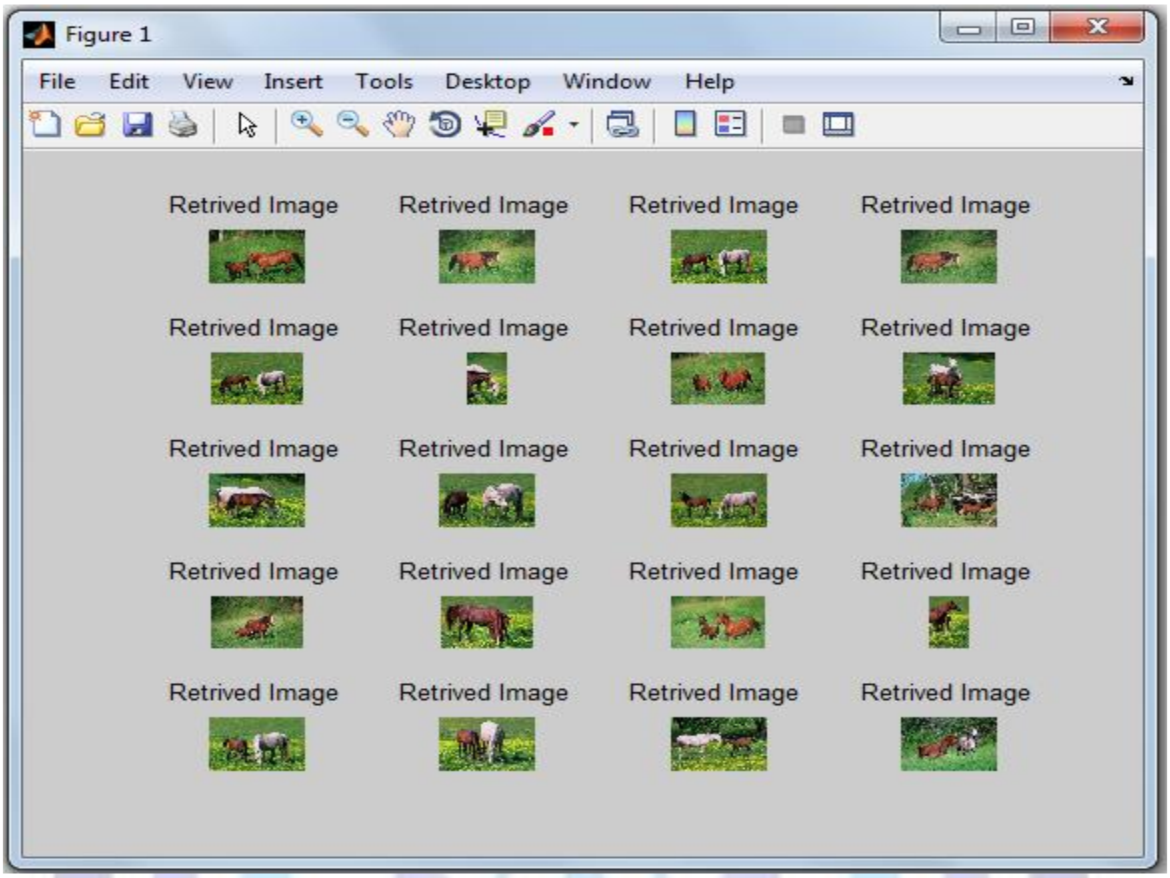


Figure 7.2: Retrieved results for Query Image of horse for proposed Image Retrieval System.

Similarly we carried out the experiments on Bus and Rose Image.We got the following results of Precision, Recall and retrieval time.We compared our results with Color Histogram method.Results are compared which is shown in table 1 below.Table 1 summarizes the precision and recall value and also the retrieval time for the results obtained for query image of Bus, Horse and Rose for our system and also of color histogram method. From the results shown, we conclude that we have substantially increased the retrieval speed but needs to be improved on retrieval results.

Table 1: Precision, Recall and retrieval time Results for Bus, Horse and Rose example.

Image Category	Method	Precision	Recall	Normal Retrieval time	Our approach (Retrieval Time)
BUS	Color Histogram	0.15	0.03	0.0322	NIL
	Our Approach	0.5	0.1	0.082	0.0242
Horse	Color Histogram	0.95	0.18	0.029	NIL
	Our Approach	1	0.19l	0.081	0.0193
Rose	Color Histogram	0.8	0.16	0.029	NIL
	Our Approach	0.95	0.19	0.064	0.0085

8. CONCLUSION

In this system, the dual problem of retrieval speed and efficiency for image retrieval system is addressed. As a result of this work conclusion can be given as:

1. In our image retrieval system by preprocessing the database we have increased the retrieval speed .i.e. images are retrieved in lesser time.

2. We have developed effective and efficient image retrieval system.

The areas for future research include image retrieval system needs to be improved on retrieval results. We can improve this by using other feature extraction methods.

REFERENCES

[1] Cai-Yun Zhao, Bian-Xia Shi , Ming-Xin Zhang, Zhao-Wei Shang,” Image Retrieval Based On Improved Hierarchical Clustering Algorithm”, Proceedings Of The 2010 International Conference On Wavelet Analysis And Pattern Recognition, Qingdao, 11-14 July 2010.

[2] Manimala Singha, K.Hemachandran,” Content Based Image Retrieval using Color and Texture”, Signal & Image Processing: An International Journal (SIPIJ) Vol.3, No.1, February 2012.

[3] Rahul Mehta, Nishchol Mishra, Sanjeev Sharma,” Color - Texture Based Image Retrieval System”, International Journal of Computer Applications (0975 – 8887) Volume 24– No.5, June 2011.

[4] Ryszard S. Chora´,”Image Feature Extraction Techniques and Their Applications for Cbir and Biometrics Systems”, International Journal of Biology and Biomedical Engineering, Issue 1, Vol. 1, 2007.

[5] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng,” Fundamentals of Content-Based Image Retrieval.”

[6] Radomir S. Stankovi_C A, Bogdan J. Falkowski B,” The Haar Wavelet Transform: Its Status and Achievements,” Computers and Electrical Engineering 29 (2003) 25–44.

[7] S.M. Zakariya, Rashid Ali and Nesar Ahmad,” Combining Visual Features of an Image at Different Precision Value of Unsupervised Content Based Image Retrieval”, 978-1-4244-5967-4/10, 2010 IEEE.

[8] http://wang.ist.psu.edu/docs/related/.

[9] Han lihua, Wang xuejun,” Design and Implementation of Image Retrieval System for Science and Technology Resources Database Based on Web”, The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010.

[10] Timo Ojala, Matti Pietikainen,” Unsupervised texture segmentation using feature distributions”, Pattern Recognition 32 (1999) 477-486.

[11] Dr.N.Rajalingam, and K.Ranjini,” Hierarchical Clustering Algorithm - A Comparative Study,” International Journal of Computer Applications (0975 – 8887) Volume 19– No.3, April 2011.

[12] A.K. Jain, M.N. Murty, And P.J. Flynn,” Data Clustering: A Review,” ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[13] Shi Na, Liu Xumin, Guan Yong,” Research On K-Means Clustering Algorithm,” IEEE Trans, Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, 978-0-7695-4020-7/10.

[14] Ruziana Mohamad Rasli, T Zalizam T Muda, Yuhanis Yusof, Juhaida Abu Bakar,” Comparative Analysis of Content Based Image Retrieval Technique using Color Histogram. A Case Study of GLCM and K-Means Clustering”, 2012 Third International Conference on Intelligent Systems Modelling and Simulation, 978-0-7695-4668-1/12, 2012 IEEE.

[15] Simona E. Grigorescu, Nicolai Petkov, And Peter Kruizinga,” Comparison Of Texture Features Based On Gabor Filters”, IEEE Transactions On Image Processing, Vol. 11, No. 10, October 2002.