Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.57 – 65

RESEARCH ARTICLE

A Novel Approach of Color Histogram Based Image Search/Retrieval

APURVA S. GOMASHE[1], PROF. R. R. KEOLE[2]

[1]Department of Comp.Sci & Engg, HVPM's COET Amravati University, Maharashtra, India
[2]Department of Information Technology, HVPM's COET Amravati University, Maharashtra, India

Abstract—Image Retrieval system is an effective and efficient tool for managing large image databases. A content based image retrieval system allows the user to present a query image in order to retrieve images stored in the database according to their similarity to the query image. The increased need of content based image retrieval technique can be found in a number of different domains such as Data Mining, Education, Medical Imaging, Crime Prevention, Weather forecasting, Remote Sensing and Management of Earth Resources. Content-based image retrieval (CBIR) scheme searches the most-similar images of a query image that involves in comparing the feature vectors of all the images in the database with that of the query image using some pre-selected similarity measure, and then sorting of the results. In this paper we see that how our new image searching/retrieving approach works and what would be the results are gives to users.

Keywords— " Image Retrieval", "Query Image", "CBIR", "Semantic gap", "Color Histogram", "Feature vector".

I. INTRODUCTION

In Web Search applications, users submit queries (i.e., some keywords) to search engines to represent their search goals. However, in many cases, queries may not exactly represent what they want since the keywords may be polysemous or cover a broad topic and users tend to formulate short queries rather than to take the trouble of constructing long and carefully stated ones. Besides, even for the same query, users may have different search goals. In various computer vision applications widely used is the process of retrieving desired images from a large collection on the basis of features that can be automatically extracted from the images themselves. These systems called CBIR (Content- Based Image Retrieval). The idea of text-based approach was originated at 1970s.In this approach the images are manually annotated by text descriptors, which are then used by a database management system (DBMS) to perform image retrieval. It has lead to two disadvantages. First one is that a considerable level of human labor is required for manual annotation. The second is the annotation inaccuracy due to the subjectivity of human perception. To overcome the above disadvantages in text-based retrieval system, content based image retrieval (CBIR) was introduced in the early 1980s. In CBIR, images are indexed by their visual content, such as color, texture, shapes. The CBIR mainly consists of two steps. One is the feature extraction and another one is the similarity matching. In various paper authors have used different feature extraction technique depending upon the low level feature or high level feature. The difference between the user's information need and the image representation is called the semantic gap in CBIR System. The system is said to be efficient if this semantic gap is minimum. With the increasing popularity of image management tools such as Google's image search and photo album tools such as Google's Picasa project, as well as image search applications in general social networking environ-ment, the quest for practical, effective image search in the web context becomes ever more important. The research community has seen a number of algorithms and tools that facilitate image retrieval. Image Retrieval aims to provide an effective and efficient tool for managing large image databases. There is a significant amount of increase in the use of medical images in clinical medicine and disease research. Image retrieval (IR) is one of the most exciting and fastest growing research areas in the field of medical imaging. The goal of CBIR is to retrieve

images from a database that are similar to an image placed as a query. In CBIR, for each image in the database, features are extracted and compared to the features of the query image. A CBIR method typically converts an image into a feature vector representation and matches with the images in the database to find out the most similar images**.** Histogram-based search methods are used in two different color spaces. Color space is defined as a model for representing color in Histogram-Based Color Image Retrieval Terms of intensity values. Typically, a color space defines a one- to four dimensional space. A color component, or a color channel, is one of the dimensions. Color spaces are related to each other by mathematical formulas. Only two three dimensional color spaces, RGB and HSV, are used. Histogram search characterizes an image by its color distribution, or a histogram. Many histogram distances have been used to define the similarity of two color histogram representations. Euclidean distance and its variations are the most commonly used. The drawback of a local histogram representation is that information about image shape, and texture is discarded. The local color histogram indexing method, which is used in this paper, correlates to the image semantics well. But, images retrieved by using the local color histogram may not be semantically related even though they share similar color distribution. Both the RGB and HSV Color spaces define a method.
In this paper is organized as follows. Section II Literature Review, In Section III Proposed Method, Section IV discuss the result analysis, Section V discuss the conclusion.

II. LITERATURE REVIEW

Currently the most popular search engines for images rely on the comparison of metadata or textual tags associated with the images. This methodology relies on human intervention to provide an interpretation of the image content so as to produce tags associated with the image. However, the ever increasing prevalence of large image databases has resulted in the development of algorithms to augment and replace tag based image retrieval with content based image retrieval. These algorithms compare the actual content of the images rather than text which has been annotated previously by a human being. There are a number of features that can be extracted from an image for comparisons based on their content. Indeed, the Photo book application developed at MIT allows users to perform image retrievals based on user developed models for various information extractions. The user specifies and provides algorithms for extracting certain features of an image in order that they be compared using the platform provided by Photo-Book. Once the specified feature has been extracted from the image, there are also a number of options for carrying out the actual comparison between images. Generally similarity between two images is based on a computation involving the Euclidean distance or histogram intersection between the respective extracted features of two images. Both these methods involve an intuitive extension of the mathematical definition of a distance between two objects. The three most common characteristics upon which images are compared in content based image retrieval algorithms are color, shape and texture. Algorithms for extracting certain features of an image in order that they be compared using the platform provided by Photo-Book. Once the specified feature has been extracted from the image, there are also a number of options for carrying out the actual comparison between images. Generally similarity between two images is based on a computation involving the Euclidean distance or histogram intersection between the respective extracted features of two images.

Lin et al. [8] proposed a color-texture and color-histogram based image retrieval system (CTCHIR). They proposed (1) three image features, based on color, texture and color distribution, as color co-occurrence matrix (CCM), difference between pixels of scan pattern (DBPSP) and color histogram for K-mean (CHKM) respectively and (2) a method for image retrieval by integrating CCM, DBPSP and CHKM to enhance image detection rate and simplify computation of image retrieval. From the experimental results they found that, their proposed method outperforms the Jhanwar et al. [5] and Hung and Dai [6] methods. Raghupathi et al. [7] have made a comparative study on image retrieval techniques, using different feature extraction methods like color histogram, Gabor Transform, color histogram+gabour transform, Contourlet Transform and color histogram+contourlet transform. Hiremath and Pujari [9] proposed CBIR system based on the color, texture and shape features by partitioning the image into tiles. The features computed on tiles serve as local descriptors of color and texture features. The color and texture analysis are analyzed by using two level grid frameworks and the shape feature is used by using Gradient Vector Flow. The comparison of experimental result of proposed method with other system [10] found that, their proposed retrieval system gives better performance than the others. Rao et al. [11] proposed CTDCIRS (color-texture and dominant color based image retrieval system), they integrated three features like Motif cooccurrence matrix (MCM) and difference between pixels of scan pattern (DBPSP) which describes the texture features and dynamic dominant color (DDC) to extract color feature. They compared their results with the work of Jhanwar et al. [5] and Hung and Dai [6] and found that their method gives better retrieval results than others.

J. Carbonell and J. Goldstein, ACM [1] introduce the Maximal Marginal Relevance (MMR) criterion strives to reduce redundancy while maintaining query relevance in re-ranking retrieved documents but the limitation of this technique is MMR is not efficient to minimize redundancy when several documents are collected.

Zhangxu-bo, IEEE [2] improved K-means clustering and relevance feedback to re-rank the search result in order to remedy the rank inversion problem in content based image retrieval but the limitation is when the visual information is totally unreliable then the K- means algorithm fail.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$Vi = (1/Ci) \sum_{j=1}^{Ci} Xi$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

Disadvantages

1) Applicable only when mean is defined i.e. fails for categorical data.
2) Unable to handle noisy data and outliers.
3) Algorithm fails for non-linear data set.

Xiaoou Tang, Fang Wen, IEEE [3] proposed a novel Internet image search approach. It only requires the user to click on one query image with minimum effort and images from a pool retrieved by text-based search are re-ranked based on both visual and textual content. In this paper limitation is its retrieve the data from only database so semantic gap is cause the problem.

Zheng Lu, Xiaokang Yang, Senior Member, IEEE [4] introduced new approach using guidance of implicit user. It designs new method which searches the images only from log history to reduced semantic gap using spectral clustering with ε-neighborhood graph. But in this system there are limitations such as the e ε-neighborhood graph is usually considered as an unweighted graph and it's fail for weighted graph. It retrieves the data only from query log if users want images that not percent in the log history then that time this system was fail.

III. PROPOSED METHODOLOGY

In this paper, image retrieval is evaluated based on color histogram of images, in following way our proposed system will work and follow. Using Euclidean distance images similar to query image is retrieved. The smaller the distance is, the more similar the two images are. The performance of the retrieval system is then evaluated [2].

- **System Description**

This section describes the overall working of the system. It describes all the phases that are present in the system. The system consists of several phases defined by an example that will make it easy for the analysis of the system.

The phases are as follows:

- **Getting Query:**

In this phase user gives or submit the query means user gives the input to the system in the form of image. In the proposed work the user has to select the image from the database and submit to the system on the basis of which relevant image in extracted.

- **Extracting Features:**

After getting a query it's important to extract the features of images, because using these extracting features of input image match with this to our existing images features. 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. Features Extraction process is play main role in every type of image search engines because its help at the time of creation of database means adding the entries into the database as well as at the time of searching images its help to find or search matching features from the database.

We classify the various features currently employed as follows:

• General features: Application independent features such as color, texture, and shape. According to the abstraction level, they can be further divided into:

- Pixel-level features: Features calculated at each pixel, e.g. color, location.
- Local features: Features calculated over the results of subdivision of the image band on image segmentation or edge detection.
- Global features: Features calculated over the entire image or just regular sub-area of an image.

• Domain-specific features: Application dependent features such as human faces, fingerprints, and conceptual features. These features are often a synthesis of low-level features for a specific domain.

On the other hand, all features can be coarsely classified into low-level features and high level features. Low-level features can be extracted directed from the original images, whereas high-level feature extraction must be based on low-level features

- **Color Features:**

The color feature is one of the most widely used visual features in image retrieval. Images characterized by color features have many advantages:

• Robustness. The color histogram is invariant to rotation of the image on the view axis, and changes in small steps when rotated otherwise or scaled [15]. It is also insensitive to changes in image and histogram resolution and occlusion.

• Effectiveness. There is high percentage of relevance between the query image and the extracted matching images.

• Implementation simplicity. The construction of the color histogram is a straightforward process, including scanning the image, assigning color values to the resolution of the histogram, and building the histogram using color components as indices.

• Computational simplicity. The histogram computation has O(X, Y) complexity for images of size X × Y . The complexity for a single image match is linear, O(n), where *n* represents the number of different colors, or resolution of the histogram.

• Low storage requirements. The color histogram size is significantly smaller than the image itself, assuming color quantization.

Typically, the color of an image is represented through some color model. There exist various color models to describe color information. A color model is specified in terms of 3-D coordinate system and a subspace within that system where each color is represented by a single point. The more commonly used color models are *RGB* (red, green, blue), *HSV* (hue, saturation, value) and Y,Cb,Cr (luminance and chrominance). Thus the color content is characterized by 3-channels from some color model. One representation of color content of the image is by using color histogram. Statistically, it denotes the joint probability of the intensities of the three color channels.

Color descriptors of images can be global or local and consist of a number of histogram descriptors and color descriptors represented by color moments, color coherence vectors or color correlogram [9].

Color histogram describes the distribution of colors within a whole or within a interest region of image. The histogram is invariant to rotation, translation and scaling of an object but the histogram does not contain semantic information, and two images with similar color histograms can possess different contents.

The standard measure of similarity used for color histograms:

- ✓ A color histogram H(i) is generated for each image h in the database (feature vector),
- ✓ The histogram is *normalized* so that its sum equals unity (removes the size of the image),
- ✓ The histogram is then stored in the database,
- ✓ Now suppose we select a *model* image (the new image to match against all possible targets in the database).

- **Feature Vector:**

In pattern recognition and machine learning, a feature vector is an n-dimensional vector of numerical features that represent some object. Many algorithms in machine learning require a numerical representation of objects, since such representations facilitate processing and statistical analysis. When representing images, the feature values might correspond to the pixels of an image, when representing texts perhaps term occurrence frequencies. Feature vectors are equivalent to the vectors of explanatory variables used in statistical procedures such as linear regression. Feature vectors are often combined with weights using a dot product in order to construct a linear predictor function that is used to determine a score for making a prediction.

Feature construction is the application of a set of constructive operators to a set of existing features resulting in construction of new features. Examples of such constructive operators include checking for the equality conditions {=, ≠}, the arithmetic operators {+, −, ×, /}, the array operators {max(S), min(S), average(S)} as well as other more sophisticated operators.

- **Comparing Color Histogram:**

Color histograms are computed for each image so as to identify relative proportions of pixels within certain values. The idea is that similar images would contain similar proportions of certain colors. The method offers numerous benefits with only a few limitations. In the proposed work we are calculating the histogram of image which is provided by the user with the images present in the database. Depending on the matching of the color histogram the relevant images are retrieved from the databases.

Color feature is one of the most widely used features in image retrieval. Colors are defined on a selected color space. Variety of color spaces include, RGB, LAB, LUV, HSV (HSL), YCrCb and the huemin- max-difference (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 content-based retrieval 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 .In my paper I have used csd as color feature. The Color Structure Descriptor (CSD [11] represents an image by both the color distribution of the image or image region (similar to a color histogram) and the local spatial structure of the color. The extra spatial information makes the descriptor sensitive to

certain image features to which an ordinary color histogram is blind. 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.

- ✓ **Color Histogram**

A color histogram represents the distribution of colors in an image, through a set of bins, where each histogram bin corresponds to a color in the quantized color space. A color histogram for a given image is represented by a vector:

H = {H [0], H [1], H [2], H [3],……H [i],……, H[n] }

Where i is the color bin in the color histogram and H[i] represents the number of pixels of color I in the image, and n is the total number of bins used in color histogram. Typically, each pixel in an image will be assigned to a bin of a color histogram. Accordingly in the color histogram of an image, the value of each bin gives the number of pixels that has the same corresponding color. In order to compare images of different sizes, color histograms should be normalized. The normalized color histogram H' is given as:

H' = {H' [0], H' [1], H' [2], H' [3],……H' [i],……, H'[n] }

Where H'=$\frac{H[i]}{p}$ p is the total number of pixels of an image.

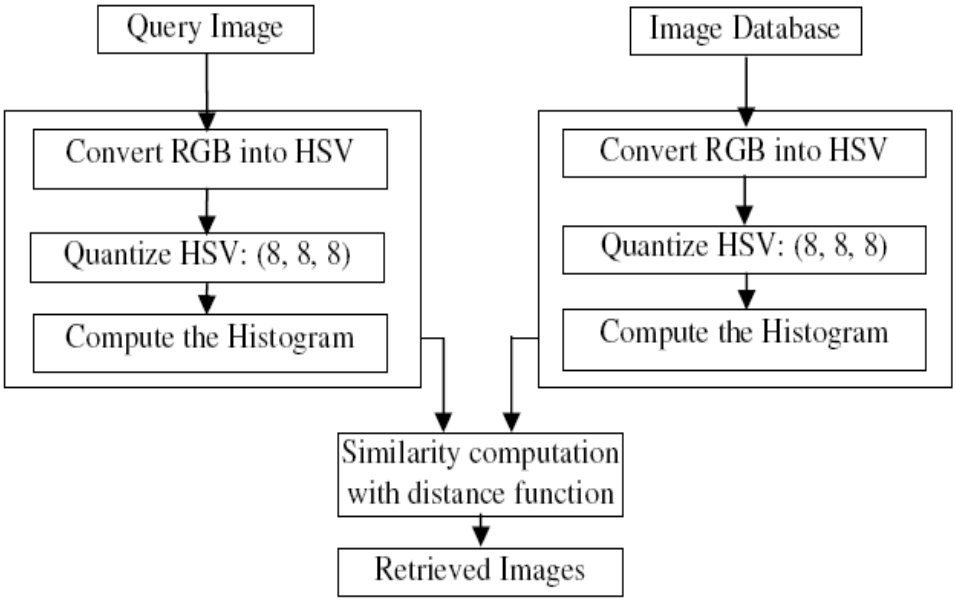

Figure 1: Block diagram of proposed Color Histogram

- **Operations perform:**

Here reranking based on Euclidean distance between query and database image is done, in which distance between input image and the relevant images in database are calculated, and the images having minimum distance is placed at the topmost position, in this way reranking is carried out.

IV. RESULT ANALYSIS

In this section we are seen how our proposed system will work with the help of screenshots in execution details and also calculate performance measurement.

- ➢ **Execution Details**

This section presents the screenshots of the Image searching/retrieving system in order to demonstrate the complete process.

1] The first screen after starting the system shown to the user is display below in the screenshots.

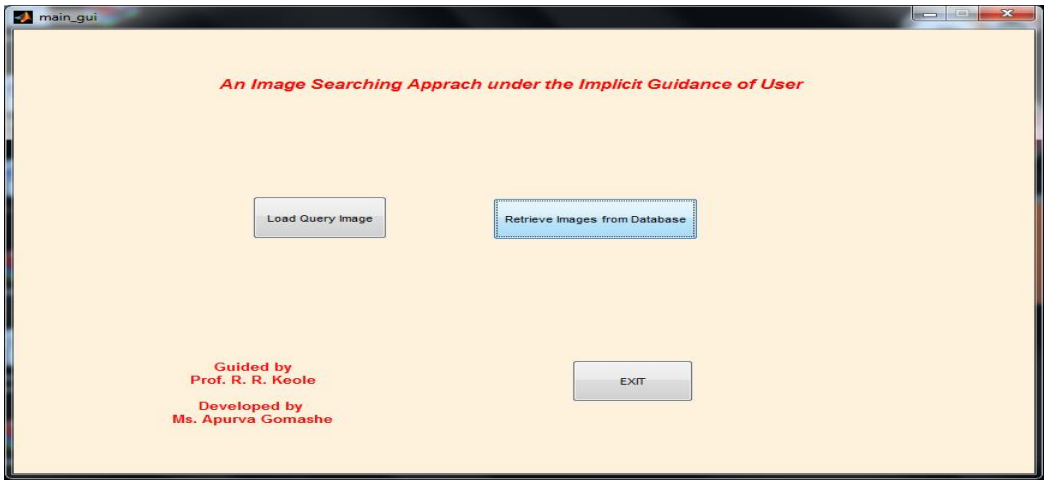

Figure 2: Home page of Image Searching and Retrieving System

Here the user has to select the images for which user wants the relevant images from the database, by clicking on the "load query image" the user provide input image to the system on which search has to be carried out.

2] Selecting the input image

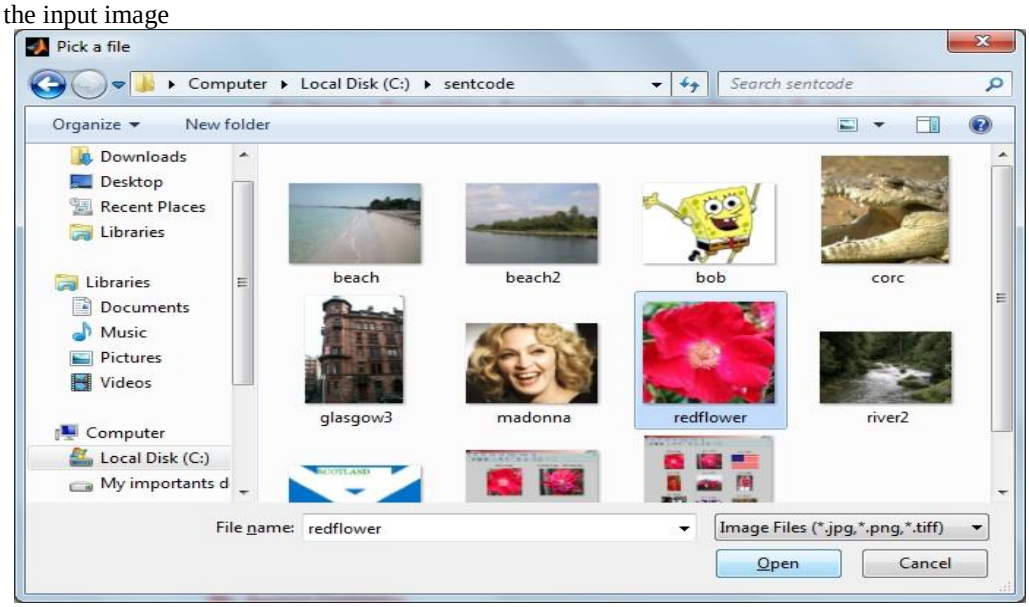

Figure 3: Input Image

Figure 3 shows the window of input image, user Browse input image from this window.

3] Number of images to be retrieved

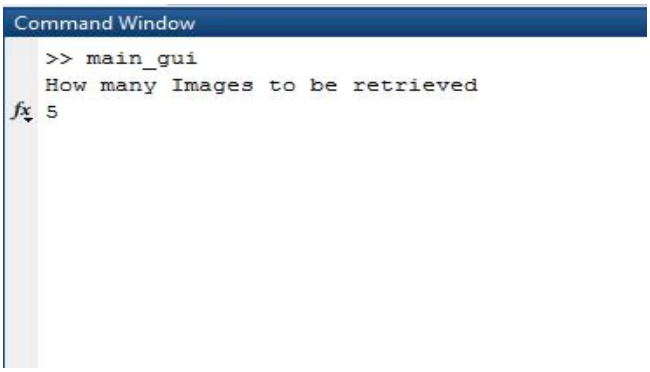

Figure 4: console to select number of images

After clicking on retrieved images from database button the system will ask user for number of images to be retrieved from the database. Depending on the number of images as input by the user, the system will find the relevant images from the database on the basis of color histogram matching. Eg: if user provides 5 number of images system will try to search and retrieve 5 numbers of relevant images from the database.

4] Calculating Color Histogram

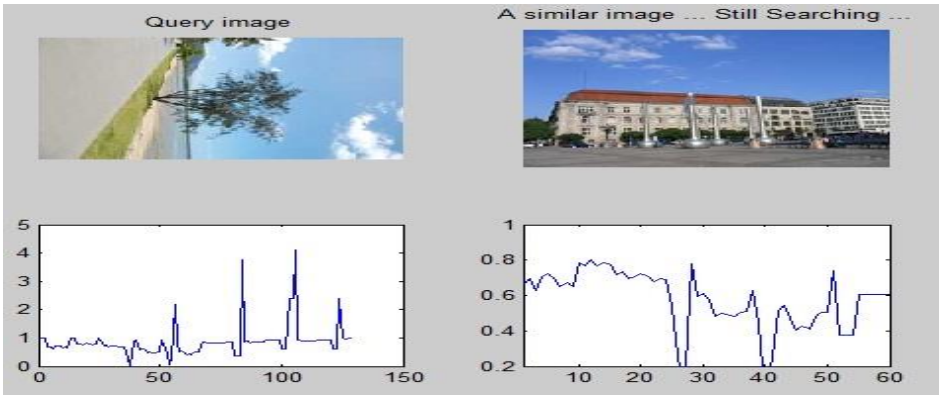

Figure 5: While the searching is being executed, some similar images are presented

Figure 5 shows calculation and matching of histogram with the image provided by the user as an input. The images from the database are compared with the images provided by the user and the relevant images are found out.

5] Resulting Relevant Images

Figure 6: When the process is completed, the query images, along with the closest images are presented.

Depending on the number of count of images provided by the user relevant images are fetched from the database, and displayed on the screen as shown in figure 6.

- **Performance Measures**

Every searcher hopes they don't retrieve a lot of junk. Unfortunately getting "everything" while "avoiding junk" is difficult if not possible to accomplish. However, it is possible to measure how well a search performed with respect to these two parameters.

Precision and recall are the basic measures used in evaluating search strategies; there is a set of records in the database which is relevant to the search. Records are assumed to be either relevant or irrelevant. The actual retrieval set may not perfectly match the set o relevant records.

- **PRECISION**: It is the ratio of number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as parentage.

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}}$$

✓ **Performance of the system**

As number of samples have been taken to carry out experiments on the system.

Sr. No	Number of images to be retrieved	Precision(%)
1	5	83.33
2	8	87
3	12	75
4	14	85
5	16	68

Table 1: Performance of System

As number of samples of various queries have been taken and search their result. Following represents result which shows better precision recall values.

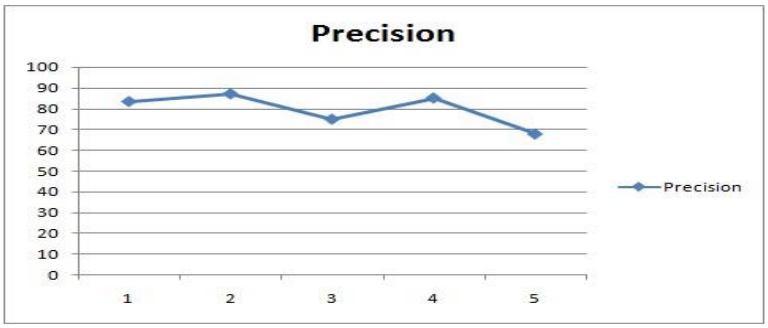

Figure7: Graph Representing Precision Vs Samples

V. CONCLUSION

By using this project images search can be done using content based, color has been taken as the property for searching, for efficient way of searching local histogram searching has been used, so it has advantages than global histogram, Euclidean distance formula has been used for comparing the histograms of the images, considering all local histograms comparisons, a sorted order of best suitable images will be generated, final search result will be displayed from that sorted order, the efficiency of this system is also obtained by calculating precession

REFERENCES

[1] Borlund, P., P. Ingwersen. The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. – Journal of Documentation, Vol. **53**, 1997, No 3, 225-250.

[2] K u i l e n b u r g, H., M. W i e r i n g, M. U y l. Model Based Methods for Automatic Analysis of Face Images. – In: Proc. of 16th European Conference on Machine Learning, 2005, 194-205.

[3] Sandeep, K., A. N. Rajagopalan. Human Face Detection in Cluttered Color Images Using Skin Color and Edge Information. – In: Proc. of Indian Conference on Computer Vision, Graphics and Image Processing, 2002.

[4] Deb, Sagarmay, and Zhang, Yanchun. "An Overview of Content-based Image Retrieval Techniques", IEEE, 18th International Conference on Advanced Information Networking and Application (AINA"04), 2004.

[5] N. Jhanwar, S. Chaudhurib, G. Seetharamanc and B. Zavidovique, "Content based image retrieval using motif co-occurrence matrix", Image and Vision Computing, Vol.22, pp-1211–1220, 2004.

[6] P.W. Huang and S.K. Dai, "Image retrieval by texture similarity", Pattern Recognition, Vol. 36, pp- 665–679, 2003.

[7] G. Raghupathi, R.S. Anand, and M.L Dewal, "Color and Texture Features for content Based image retrieval", Second International conference on multimedia and content based image retrieval, July-21- 23, 2010.

[8] C.H. Lin, R.T. Chen and Y.K. Chan, "A smart content-based image retrieval system based on color and texture feature", Image and Vision Computing vol.27, pp.658–665, 2009.

[9] P. S. Hiremath and J. Pujari, "Content Based Image Retrieval based on Color, Texture and Shape features using Image and its complement", 15th International Conference on Advance Computing and Communications. IEEE. 2007.

[10] J. Li, J.Z. Wang and G. Wiederhold, "IRM: Integrated Region Matching for Image Retrieval", In Proceeding of the 8th ACM International Conference on Multimedia, pp- 147-156, Oct. 2000.

[11] M. B. Rao, B. P. Rao, and A. Govardhan, "CTDCIRS: Content based Image Retrieval System based on Dominant Color and Texture Features", International Journal of Computer Applications, Vol. 18– No.6, pp-0975-8887, 2011.

