

# Color Histogram Features for Image Retrieval Systems

Sanmukh.N.Mali, Tejaswini.M.L

P.G. Student, Dept. of E&C, Don Bosco Institute of Technology, Bangalore, Karnataka, India

Assistant Professor, Dept. of E&C, Don Bosco Institute of Technology, Bangalore, Karnataka, India

**ABSTRACT:** Histogram features have proved powerful in the classification of image and object detection . The CBIR most efficient and searches the color based images. Here in this method we use some improved preprocessing steps, preprocessing algorithms and the image classification is analyzed. In CBIR image classification has to be computationally very fast and efficient. In this project a new approach is introduced, which based on low level image histogram features. Color is a main powerful descriptor that often identifies object and extraction scene. The main advantage of this method is the very quick generation and comparison of the applied feature vectors. Histograms are simple to calculate in software and also lend themselves to economic hardware implementations. A popular tool for a real-time image processing histogram-based image retrieval methods in two color spaces were exhaustively compared. The testing also highlights the weaknesses and strengths of the model.

**KEY WORDS:** color-based and content based image retrieval, K-Nearest neighbor method, histogram intersection and its similarity.

## I. INTRODUCTION

In (CBIR) which is abbreviated as content-based image retrieval systems is very much useful and efficient if the images are classified on the score of particular aspects. 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. The use of colour in image processing is motivated. Let us consider a great database the images can be divided into such classes: trees, animals, faces, artificial images, landscapes, etc. testing also highlights the weaknesses and strengths of the model. Many color image classification methods use color histograms. The three most common characteristics upon which images are compared in content based image retrieval algorithms are color, shape and texture. where the so-called blobworld is used to search similar images. The aim of this project to develop such a color histogram based classification approach, which is efficient, quick and enough robust. In the interest of this I used some features of color histograms, and classified the images using these features. The advantage of this approach is the comparison of histogram features is much faster and more efficient than of other commonly used methods. Firstly, image retrieval based on this concept should accurately retrieve images despite the manipulation of orientation, size and position of a certain image.

## II. LITERATURE REVIEWS

A most basic form of color retrieval involves specifying color values that can be searched for in images from a database .let us consider recognition is an easy task for humans. That even two or three days old babies are able to distinguish between known faces .So how hard could it be for a computer? It turns out we know little about human recognition to date. are inner features that are eye, nose, mouth and the outer features are head shape , hairline, body shape used for successful humans recognition .

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, March 2014

By Referring C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, J. Malik, “Blobworld [1] I got an concept of a System for Region-Based on Image Indexing and Retrieval,” Third International Conference on Visual Information of Systems Here I got information on the systems for retrieval based on finding coherent image regions which roughly correspond on object but it rather description of entire image, from this experiment I learnt resulting for both querying and indexing.

According to the concept of Sz. Sergyan [2] , “Color Image Classification”2000 , 5th Slovakian- Hungarian Joint Symposium on Applied Intelligence and Informatics I got to know that the image systems are most efficient and simple searches and are color based searches, which can be realized in several color spaces and by several color spaces and the several color descriptors and I also got to know that classification using the certain color descriptors and usage of different color spaces depending upon image database doma in is presented.

With the help of A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain [3], “ Image based Retrieval” 2000, I came to know about the computational steps of image retrieval system like the color ,texture and local geometry. Then accumulative and global features, salient points ,object and shape features, signs and structural combinations thereof. In concluding section he presents on the driving force of field, the heritage from computer vision,and the influence on computer vision,the role of interaction.

### III EXTRACTION OF CANDIDATEBUILT-UP REGIONS

#### A. Color Based Retrieval

A number of algorithms have been developed since the late 1980s that use color information extracted from images for retrievals. Here histogram features that we will consider are data based features, where the histogram is used as a model of the probability distribution of the intensity levels Most intuitive information that can be extracted from images for comparison is the color characteristics of an image. Here we explore and analyze such an algorithm that compares images based on their color content according to database. These database features provides us with information about the characteristics of the intensity for distribution for the image. We define the histogram probability,  $P(g)$  as:

$$P(g) = N(g) / M$$

Where  $M$  is number of pixels in the image (if the entire image is under consideration then  $M = N^2$  for an  $N \times N$  image), and  $N(g)$  is the number of pixels at gray level  $g$ . As with any probability distribution all the values for  $P(g)$  are less than or equal to 1, and the sum of all the  $P(g)$  values is equal to 1. The features based on the first order histogram probability are the mean, standard deviation, skew, energy, and entropy. The mean is the average value, so it tells us something about the general brightness of the image. A bright image will have a high mean, and a dark image will have a low mean. We will use  $L$  as the total number of intensity levels available, so the gray levels range from 0 to  $L - 1$ . For example, for typical 8-bit image data,  $L$  is 256 and ranges from 0 to 255. We can define the mean as follows:

$$\sum_{g=0}^{L-1} gP(g) = \sum_c \sum_r I(r, c) / M$$

If we use the second form of the equation we sum over the rows and columns corresponding to the pixels in the image under consideration. The standard deviation, which is also known as the square root of the variance, tells us something about the contrast. . The standard deviation, which is also known as the square root of the variance, tells us something about the contrast. It is defined as follows:The skew measures the asymmetry about the mean in the intensity level distribution. It is defined as:

$$SKEW = \frac{1}{\sigma_g^3} \sum_{g=0}^{L-1} (g - \bar{g})^3 P(g)$$

The skew will be positive if the tail of the histogram spreads to the right (positive), and negative if the tail of the histogram spreads to the left (negative). Another method to measure the skew uses the mean, mode, and standard deviation, where the mode is defined as the peak, or highest, value. Most image database systems are products of

## International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, March 2014

research, and therefore emphasize only one aspect of content-based retrieval. Sometimes this is the sketching capability in the user interface; sometimes it is a new indexing data structure, etc. Some systems are created as a research version and a commercial product. The commercial version is usually less advanced, and shows more standard searching capabilities. A number of systems provide user interface that allows more powerful query formulation than is useful in demo system. Most systems use color and texture features, few systems use shape features, and yet less use spatial features. The retrieval on color usually yield images that have similar colors. The larger the collection of images, the greater is the chance that it contains an image similar to the query image. This method of measuring skew is more computationally efficient, especially considering that, typically, the mean and standard deviation have already been calculated. It describes the spread in the data, so a high contrast image will have a high variance, and a low contrast image will have a low variance. The energy measure tells us something about how the intensity levels are distributed:

$$\text{ENERGY} = \sum_{g=0}^{L-1} (P(g))^2$$

The energy measure has a maximum value of 1 for an image with a constant value, and gets increasingly smaller as the pixel values are distributed across more intensity level values (remember all the  $P(g)$  values are less than or equal to 1). Finally I summarize the theoretical background of my classification method. The certain subsections are based on. The histogram of an image is a plot of the gray level values or the intensity values of a color channel versus the number of pixels at that value. The shape of the histogram provides us with information about the nature of the image, or subimage if we are considering an object within the image. For example, a very narrow histogram implies a low contrast image, the histogram skewed toward the high end implies a bright image, and the histogram with two major peaks, called bimodal, implies an object that is in contrast with the background. The histogram features that we will consider are statistical based features, where the histogram is used as a model of the probability distribution of the intensity levels. These statistical features provides us with information about the characteristics of the intensity level distribution for the image. We define the first-order histogram probability. As the pixel is on values in the image are distributed among more intensity levels, the entropy increases. The features based on the first order histogram probability are the mean, standard deviation, skew, energy, and entropy. The mean is the average value, so it tells us something about the general brightness of the image. A bright image will have a high mean, and a dark image will have a low mean. We will use  $L$  as the total number of intensity levels available, so the gray levels range from 0 to  $L - 1$ . For example, for typical 8-bit image data,  $L$  is 256 and ranges from 0 to 255.

### B. Classification Algorithm and Methods

The simplest algorithm for identifying a sample test set is called the Nearest Neighbor method. Here we compare the training set, using a distance and similarity measure, or a combination of measures. For the unknown object is then identified as belonging to the same class as the closest sample in the training set.

Nearest Neighbor method is more robust by selecting not just the closest sample in the training set, but by consideration of a group of close feature vectors. This is called the K-Nearest Neighbor method, let use consider example,  $K = 5$ . Then we assign the unknown feature vector to the class that occurs most often in the set of K-Neighbors. This is still very computationally intensive, since we have to compare each unknown sample to every sample in the training set, and we want the training set as large as possible to maximize success. Computational burden is reduced by the method called Nearest Centroid. Here, we find the centroids for each class from the samples in the training set, and then we compare the unknown samples to the representative centroids only. The motivation behind is this is a new approach for the feature vector is meant to represent the object and will be used to classify it. To perform the classification we need methods to compare two feature vectors. The primary methods are to either measure the difference between the two, or to measure the similarity. Then two vectors that are closely related will have a small difference and a large similarity. The difference can be measured by a distance measure in the  $n$ -dimensional feature space. The bigger the distance between two vectors, the greater the difference. Euclidean distance is the most common metric for measuring the distance between two vectors.

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, March 2014

## C. Feature Vectors and Its Size

The feature vector is an n-dimensional vector that contains these measurements, where n is the number of features. The measurements may be symbolic, numerical, of both For example with color we normally think of the hue by its name such as "orange" or "magenta". In this case, we could perform an HSL transform on the RGB data, and use the H (hue) value as a numerical color feature. But with the HSL transform the hue value ranges from 0 to 360 degrees, and 0 is "next to" 360, so it would be invalid to compare two colors by simply subtracting the two hue values.

The feature vector can be used to classify an object, or provide us with condensed higher-level image information. Associated with the feature vector is a mathematical abstraction called a feature space For n-dimensional feature vectors it is an abstract mathematical construction called a hyperspace.

## IV. IMAGE COLLECTION AND EXPERIMENT SET-UP

The experiments uses 200 several images, which was divided into four equal size classes: landscapes, buildings, faces and indoor images with one object with homogenous background. One image on each classes are shown in the below figure.

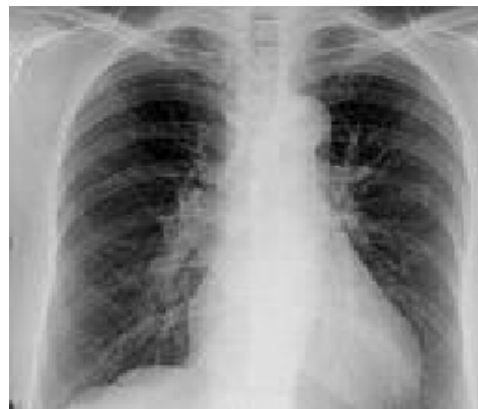


Fig.1. Sample images of each class: landscape, building, face, indoor image with one object

## International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, March 2014

From each image classes 25 images were the member of the training class. Using each training set the histograms of the three color channels were generated, and the above mentioned histogram features were calculated. Hence in each training set there were 25 pieces and 15-dimensional feature vectors and are madeup of 15-dimensional hyperspace. After the property generation of the training set I analyzed that the remaining 100 images are closest to which class. I found the 87% of images were well classified during the experiment. Here the algorithm is coded by MATLAB because here system is computationally faster and code generator is very simple, and one such example for MATLAB code for histogram feature generator can be seen here below :

```
function t = HistogramProperties ( PrH, N)
t . m = sum([1:N]' .* PrH);
t.s = sqrt (sum((([1:N]'-t.m).^2.*PrH));
t.sk1= sum ((([1:N]'-t.m).^3.* PrH)/t.s.^3);
mode = find(PrH == max(PrH));
t.sk2 = (t.m-mode(ceil(length(mode)/2)))/t.s;
t.e r= sum(PrH.^2);
t. ep = -sum(PrH .* log2(PrH + eps));
```

Fig.2. MATLAB code of the generation of histogram features

### APPLICATIONS

- Body part recognition, object recognition in medical applications.
- Detecting Man-Made Structures and Changes in Satellite Imagery.
- Resolution Remote Sensing Images
- Image classification
- Camera capture image.

### ADVANTAGES

- Usage of simple image features, as histogram features.
- Histogram features can be Very Quickly and the Comparison of computationally fast and efficient.
- More image classes and more than thousand images.
- Very speed comparison of colored images.
- Used in Hospital applications, library application, Industrial applications
- Very high speed in identifying the difference of image with respect to color.

### DISADVANTAGES

- It can not recognize unstored data.



Figure2 output image



# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, March 2014

## V. CONCLUSIONS

The new approach in my presentation is of color image classification is introduced and advantage is usage of simple image features and histogram features. Histogram can be generated from the image histogram very quickly and the comparison of this features is computationally fast and efficient. In further works a bigger test seems very important. I will make similar experiment with more image classes and more than thousand images.

## REFERENCES

- 1) C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, J. Malik, "Blobworld: A System for Region-Based image Indexing and Retrieval," Third International Conference on Visual Information Systems, Springer, 1999.
- 2) AIRS- Advanced Image Retrieval <http://www.imageclick.com/airs/sub/aboutAIRs.html> (accessed October 2008).
- 3) Content-based image retrieval – Wikipedia Nov 2008;<http://en.wikipedia.org/wiki/CBIR> (accessed November 8, 2008).
- 4) Sz. Sergy'an, "Color Content-based Image Classification," 5th Slovakian- Hungarian Joint Symposium on Applied Machine Intelligence and Informatics, Poprad, Slovakia, pp. 427–434, 2007.
- 5) A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content- Based Image Retrieval at the End of the Early Years," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349–1380, 2000.
- 6) S.E. Umbaugh, "Computer Imaging – Digital Image Analysis and Processing," CRC Press, 2005.