Image Clustering using Color Moments, Histogram, Edge and K-means Clustering

Annesha Malakar[1], Joydeep Mukherjee[2]

[1]Jadavpur University, School of Education Technology,
Kolkata, India
annesha.sonu@gmail.com

[2]Jadavpur University, School of Education Technology,
Kolkata, India
joym761@email.com

Abstract: ***Clustering a large volume of image database is a challenging research work. Image clustering is needed many practical area like Medical Diagnosis, Military. There exist many traditional way to cluster similar data. But the accuracy level is not so high. So in this paper we propose a new multi feature image clustering technique which will help us to classify the large volume data with high accuracy level. Firstly we extract color moments feature from an image, and then we consider histogram analysis and make a summation of each color bin. Finally we used canny edge detection technique. Lastly we combine all features in a matrix and perform clustering algorithm to cluster data.***

Keywords: Color moments, Color Histogram, Edge Detection, Clustering.

1. Introduction

With the popularity of computer based system, CBIR takes a great challenge of research. Everywhere we see the usage of image and image plays a vital role in different area e.g. Medical Diagnosis, Military, Retail Catalogs etc. Therefore we need an efficient technique to retrieve such images. There exist many traditional information retrieval techniques but they do not meet the user's demand with increasing the volume of image data.

There are many way to retrieve an image. Feature extraction is the key function of any Content Based Image Retrieval System. Feature extraction means mapping the image pixels into the feature space. Using this extracted feature we can search, indexed and browse the image form the stored database and this feature can be used to measure the similarity between the stored images.

The feature of image can be classified into middle level feature and low- level feature. Low-level feature includes color, texture and inflexion. Middle level involves shape description and object feature. Among these color feature can be extracted in many ways like Histogram [1], Color moments [2], Color Correlogram [3] etc. Texture feature can be extracted using Gray Level Co-occurrence matrix (GLCM) [4], Gabor Filter Response [4] etc.

Edge detection technique also needs to classify different images. There are many edge detection techniques like Robert, Sobel, Prewitt, Kirsch, Robinson, Marr-Hildreth, LoG and Canny Edge Detection [5]. Many researchers tells that Marr-Hildreth, LoG and Canny produce the same result where Kirsch, Robinson produce the same also. But canny gives better result than others.

In shape or texture feature has some disadvantage. When an object is rotated or scaled then it treats as a different one. But color feature does not face the problem. So we take color feature as an important role in our system.

In the paper[15] Clustering Similar Image of Remote Sensing Images is done based on Using Colour Moment Feature Detector and K- means Clustering. Now in this paper we propose an image store and clustering method based on Color Moments, Color Histogram analysis and Canny Edge Detection technique and k-means [6] technique for clustering the data.

Our first approach on feature extraction is to extract the color moments value from an image. The mean, variance and standard deviation of an image are known as color moments [7]. This value is stored in a 1D array. This value is calculated for every image in the database.

Our second approach on feature extraction is image Histogram analysis. The color histogram for an image is constructed by quantizing the colors within the image and counting the number of pixels of each color. The histogram-based method is very suitable for color image retrieval because they are invariant to geometrical information in images, such as translation and rotation. This image histogram is like a bar graph and these values are stored in the same 1D array. This value is calculated for every image in the database.

Our third approach in feature extraction is edge detection of an image. For this purpose we used Canny's Edge detection technique which gives most effective results for our proposed system. Here we get another 1D array.

Finally the entire three 1D array are merged in a single 1 D array of extracted feature for single image. This process is repeated for every image. Then we apply k- means clustering algorithm to cluster the collection of data objects that are similar to one another within the same cluster and store the dissimilar objects in the other clusters.

The rest of the paper organized as follows, section 2- Related Work surveyed by us, section 3- System Overview,

section 4- Proposed Work, section 5- Clustering ,next section is about experimental result that contains accuracy table, and this is followed by conclusion and future scope.

2. Related Work

In past years, some paper has been presented for clustering image database. Some clustering method is used to group similar data. The image type, size, color and texture characteristics are extracted from the images and stored into the database as metadata. Based on the stored data some clustering algorithm is used to group the data. But besides this offline approaches some online approaches [8] are also appreciable in this context. This approach has both higher level and lower level feature extraction. The higher level is just the refinement of lower level feature extraction. And with the introduction of finer features number of candidate images gradually decreases and search become more efficient.

Color is one of the most widely used features for image retrieval. Color is invariant to complexity and very much sensitive to humans than the grayscale images. The color features are mostly extracted using the color histogram techniques, Color coherence vector etc. The color distributions in the images are represented using the Color Histograms. The histogram of the query image and the database images are compared for the retrieval. This method fails in case two different images have the same color distributions. These color moments extracts not only the color distribution of pixels in images like color histogram, but also extracts the spatial information of pixels in the images. It gives us a more sophisticated approach towards histogram refinement. Some efficient work on color moments is done for clustering similar image [15].But color moments itself cannot give the best accuracy. So we used color moments and color histogram technique as a combined feature to get better accuracy level.

Most of the works related to the content based image retrieval is associated with the color extraction feature. If the RGB color space is used in some approach of image clustering then researcher get high priority to the red, green and blue values. And in case of HSV color space, the hue, saturation and brightness gives the high priority level .Jagadeesh Pujari, Pushpalatha S.N, Padmashree D. Desai [9] used HSV and Lab color space to recognize an image and then compared it with grey and RGB approach. In their experiment Lab color space gives better result than other ones. But, Young Deok Chun, Nam Chul Kim and Ick Hoon Jang‟s [10] proposed approach is based on the HSV color space. This hue and saturation component of an image is stored in a array. And it gave a higher accuracy level than some other conventional methods. Some approaches also used the database to sore the feature value of the images that stores color values as well as other features values. Then some clustering algorithm is used to group similar data. Xiang-Yang Wang, Yong-Jian Yu, Hong-Ying Yang [11] proposes a system that firstly clusters the image. Here the image is predetermined using fast color quantization algorithm. Then spatial texture feature is extracted from an image and merged with the previous values and finally a robust system is presented.

Shape detection of an image is an important feature for object recognition. Shape description or representation of edge is an important issue for clustering. There exists different edge detection technique to detect edges of objects in the image. Like Robert, Sobel, Prewitt, Kirsch, Robinson, Marr-Hildreth, LoG and Canny Edge Detection [12]. N. Senthilkumaran and R. Rajesh [13] have done a comparative study in different edge detection techniques. They have used the soft computing approaches namely, fuzzy based approach, Genetic algorithm based approach and Neural network based approach. In their research it is seen that Robert method is better than both Sobel and Prewitt method .But by our experiment we get the conclusion from our system that Canny method gives more accuracy than the Robert method.

3. System Overview

Our proposed methodology involves multi feature extraction method e.g. color moments, color histogram and canny edge detection technique. The overall system overview is shown in the Figure 1.

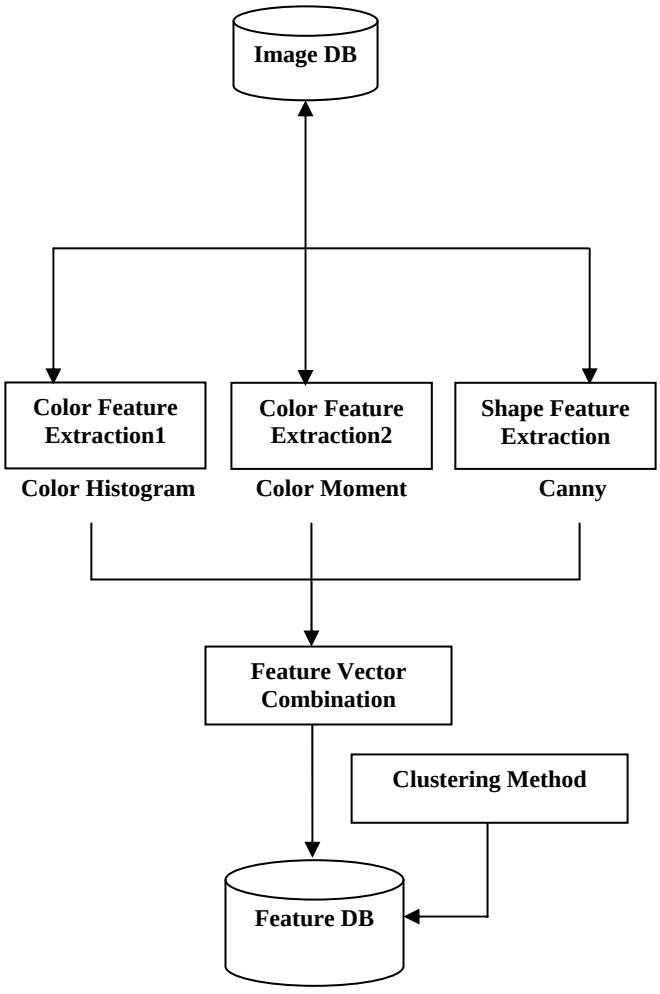

Figure 1: System Overview

3.1 Image Color Moments Analysis Method

Color moments are used to differentiate images based on their features of color. This moment is used to measure the color similarity between images. The basis of color moments lays in the assumption that the distribution of color in an image can be interpreted as a probability distribution. If the color in an image follows a certain probability distribution, the moments of that distribution can then be used as features to identify that image based on color.

Stricker and Orengo [14] use three central moments of an image's color distribution. They are Mean, Standard deviation and Skewness. A color can be defined by 3 or more values (Red, Green, and Blue).Moments are calculated for each of these channels in an image. An image therefore is characterized by 9 moments 3 moments for each 3 color channels.

We will define the i-th color channel at the j-th image pixel as *Pij*. The three color moments can then be defined as:

MOMENT 1 - Mean

MOMENT 2 - Standard Deviation

MOMENT 3 - Skewness

3.2 Image Color Histogram Analysis Method

The histogram provides a compact summarization of the distribution of data in an image. The color histogram of an image is relatively invariant with translation and rotation about the viewing axis, and varies only slowly with the angle of view. Color histogram of an image is a type of bar graph and these acts as a graphical representation of the tonal distribution in a digital image.

The number of elements in a histogram depends on the number of bits in each pixel of an image. For example, if we consider a pixel depth of n bit, the pixel values will be in between 0 and $2^n - 1$, and the histogram will have 2^n elements. In our method we consider 8 bit image. So, there are 256 different bins of color for three color space Red, Green and Blue.

In our proposed approach image histogram feature is extracted for three color space Red, green and Blue. And the histogram value matrix holds no of pixels of 256 different bins. This picture is shown in Figure 2

3.3 Edge Detection Method

All of the image pixels in an object are similar with respect to some characteristics such as color, intensity, or texture. There are different approaches to distinguish the different section of a image. Among this (i) by finding boundaries between regions based on discontinuities in intensity levels, (ii) thresholds based on the distribution of pixel properties, such as intensity values, and (iii) based on finding the regions directly. Here we used Region Based method which is based on continuity. These techniques divide the entire image into sub regions depending on some rules. Region-based techniques rely on common patterns in intensity values within a cluster of neighboring pixels. These region based methods are also called as Edge or Boundary based methods. There are various techniques of edge detection available i.e. canny method, LoG method, Sobel method, Prewitt method, Robert method etc. Here we first convert an image into gray color space and then apply the above said formula.

From the Figure 3 we see that Canny Edge Detection (Canny [1986]) [12] Technique gives best result. It was first created by John Canny for his Master's thesis at MIT in 1983, and still outperforms many of the newer algorithms that have been developed.

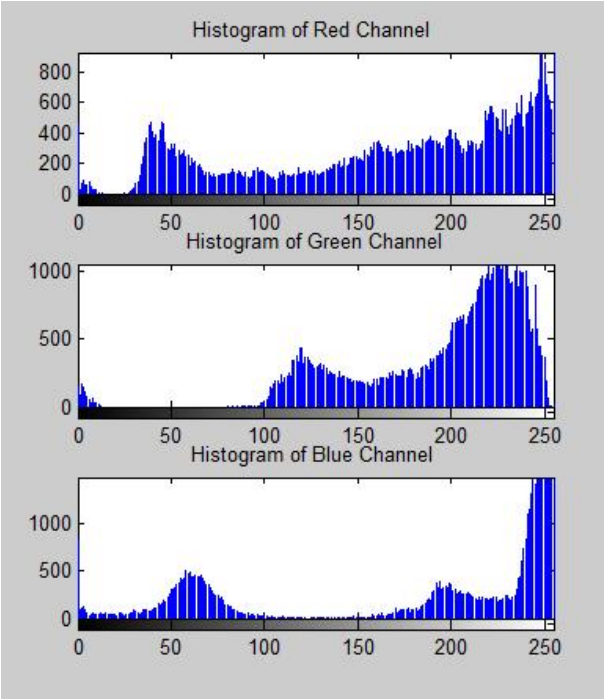

Figure 2: Image histogram analysis diagram

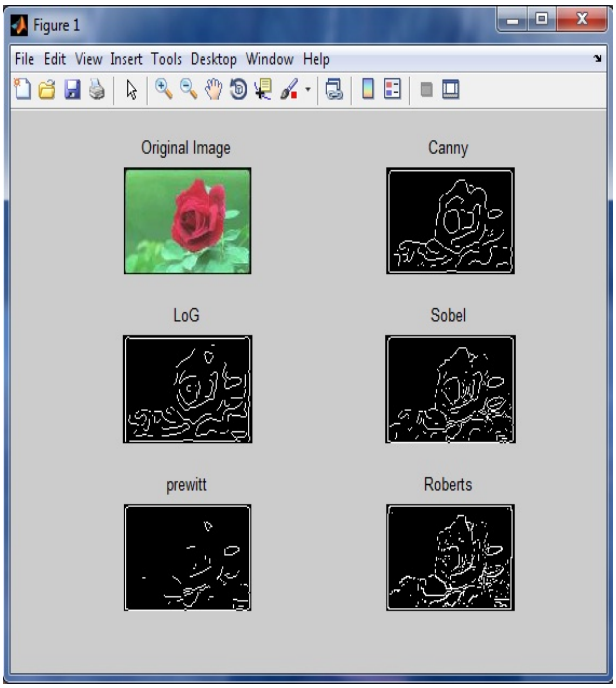

Figure 3: Image Edge detection method

4. Proposed Work

As mentioned earlier, our proposed approach uses three feature extraction methods i.e. image color moment's extraction in RGB space, image color histogram extraction in RGB space and image edge detection using Canny's edge detection technique. Here we consider 8 bit RGB image for our method.

4.1 Color Moments Feature Extraction

Step 1**:** Read the image file.

Step 2**:** We find the mean value using the following function

$$Ei = \sum_{j=1}^{N} Pij$$

Step 3**:** We find the Standard Deviation value using the following function

$$\sigma i = \sqrt{\frac{1}{N}\left(\sum_{j=1}^{N} (Pij - Ei)2\right)}$$

Step 4**:** We find the Skewness value using the following function

$$S_i = \sqrt[3]{\left(\frac{1}{N}\sum_{j=1}^{N} (P_{ij} - E_i)3\right)}$$

Step 6**:** Finally store the value of mean, standard deviation and skewness in a 1D array.

Step 7: Step 1-4 is repeated for every image in the database.

4.2 Histogram Feature Extraction

Step 1**:** Read the image file.

Step 2**:** Convert the image pixel value to a double data class.

Step 3**:** Store the value of each red, green, blue component in three different arrays.

Step 4**:** Find the image histogram of red, green, blue component by MATLAB's own histogram computational method.

Step 5: Store the histogram value of red, green, blue component in three arrays.

Step 6: Calculate the sum of 256 different bin of red component. Apply this method for green and blue component also.

Step 7: Finally merged the value of red, green, blue component in the previously created 1D array.

Step 8: Step 1-8 is repeated for every image in the database.

4.3 Edge Detection and Values Extraction

Our proposed algorithm to find the edge of a color image is as follows:

Step 1**:** Read a color image.

Step 2**:** Convert the color image into Gary Color space. Because MATLAB support edge detection technique only on Gray scale image.

Step 3**:** The image is smoothed using Gaussian Filter to reduce the Noise.

Step 4: First difference gradient operator is used to compute the edge strength and edge direction.

Step 5**:** Then Non- Maximal Suppression is performed to the gradient magnitude.

Step 6**:** Then the algorithm Performs edge links by incorporating 8-connected method which is called hysteresis operator, in which pixels are marked as either edges, non-edges and in-between, this is done based on threshold values.

Step 7**:** Finally these values gives us 2D array of edge. We take summation of the matrix by row and column wise, which gives a single value of the image. It is stored on the Featured Database.

Step 8**:** Step 1-8 is repeated for every image in the database.

5. Clustering

Clustering means to group the similar data objects that are similar to one another with in the same cluster and are dissimilar to the objects in the other clusters.

This part briefly describes the standard k-means algorithm. K-means is used to cluster data in data mining application. In 1967, MacQueen firstly proposed the k-means algorithm. It was one of the most simple, non-supervised learning algorithms, which was applied to solve the problem of the well known cluster. The algorithm consists of two separate phases.

a. In the first phase we selects k centers randomly where k is given by us.

b. In the next phase we measure each data object to the nearest center. Here Euclidian distance is used to determine the distance between each data object and the cluster center. When the entire data objects are included in any cluster, the first step is completed and we get an initial group.

c. Again the same process is done and recalculates the cluster.

d. If the group formed is same before the last iteration, then iteration reveals that object does not move anymore. Thus the computation of the k-means clustering has reached its stability and no more iteration is needed. We

get the final groping as the result. If not then again start the same process.

The Euclidean distance D(X, Y) can be obtained as follow:

$$D(X_i, Y_i) = \sqrt{\sum_{i=1}^{n}(Xi - Yi)}2$$

6. Results and Discussion

This system is implemented using MATLAB image processing tools and statistical tools. For the experiment, we take 12 images (Figure 4) and perform feature extraction followed by Clustering algorithm which gives 4 cluster of image as shown in the Figure 5.

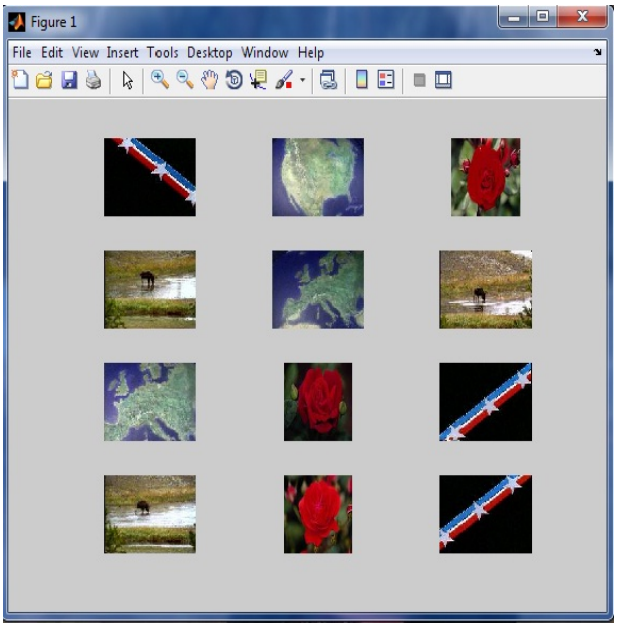

Figure 4: Result before clustering

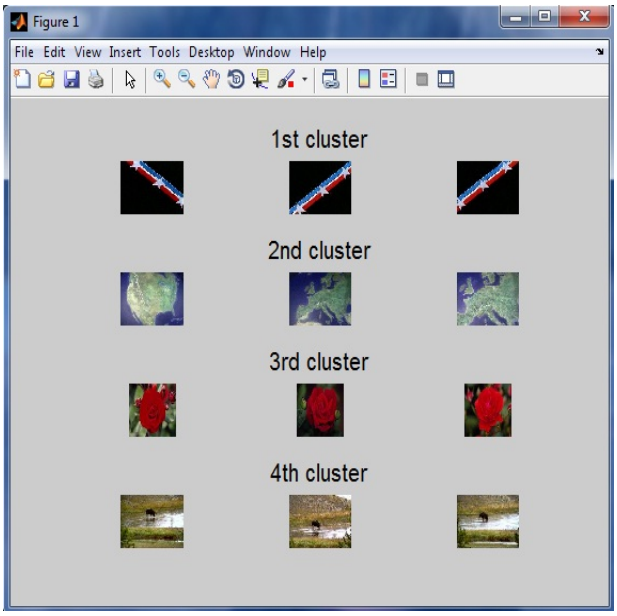

Figure 5: Result after clustering

We also check the result using 40 images and measure accuracy percentage. The Table 1 gives the result of accuracy.

Table 1: Accuracy Table

No of image	*Accuracy Percentage*
10	92%
20	95%
30	88%
40	87%
Overall accuracy	90.5%

7. Conclusions and Future Scope

In our proposed approach we worked with image retrieval from some stored images. If we can maintain proper image database, then the complexity of clustering the image will be decreases.

Euclidean distance method is used to calculate the K-means Clustering. Here Chebyshev distance, Manhattan distance or any Neural Network method can give better accuracy level.

References

[1] M L Swain,H H Ballard,"color indexing", International Journal of Computer Vision, 7(1):pp.11-32, 1991

[2] M Stricker, M Orengo, "Similarity of color images", In : W Niblack,R C Jain eds.Pro of SPIE Storage and Retrieval for Image and Video Databases, Vol 2420. San Jose, CA, USA: SPIE Press, p381-392, 1995

[3] J Huang,S R Kumar, M Mitra et al, "Image indexing using color correlograms", Proc of IEEE Conf on Computer Vision and Pattern Recognition, San Jose,Puerto Rico,USA: IEEE CS Press, p762-768,

[4] 1997

[5] D.S. Guru, Y.H. Sharath & S. Manjunath, Texture Features and KNN in Classification of Flower Images,, IJCA Special Issue, 2010.

[6] Ehsan Nadernejad. Sara Sharifzadeh &, Hamid Hassanpour, *Edge Detection Techniques: Evaluations and Comparisons.* 2nd ed., Applied Mathematical Sciences, 2008

[7] Khaled Alsabti. Sanjay Ranka &, Vineet Singh, *An Efficient K-Means Clustering Algorithm,*,

[8] Hu M.K. (1962), Visual pattern recognition by moment invariants, computer methods in image analysis. IRE transactions on Information Theory, Vol. 8

[9] Jozsef Vass, Jia Yao, Anupam Joshi, Kannappan Palaniappn, Xinhua Zhuang, "Interactive image retrieval over the inyternet", Reliable Distributed Systems, 1998. Proceedings, Seventeenth IEEE Symposium on 20-23 Oct 1998, pp: 461 – 466.

[10] Jagadeesh Pujari , Pushpalatha S.N, Padmashree D.Desai, "Content-Based Image Retrieval using Color and Shape Descriptors", Signal and Image Processing (ICSIP), 2010 International Conference on, pp: 239 – 242

[11] Young Deok Chun, Nam Chul Kim and Ick Hoon Jang, "Content-Based Image Retrieval Using Multi resolution Color and Texture Features", Multimedia, IEEE Transactions on Oct. 2008, Volume: 10, Issue: 6, pg no: 1073 – 1084.

[12] Xiang-Yang Wang, Yong-Jian Yu, Hong-Ying Yang," An effective image retrieval scheme using color, texture and shape features", Published in: · Journal Computer Standards & Interfaces archive Volume 33

Issue 1, January, 2011 Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands.
[13] Marr, D & E. Hildreth (1980) "Theory of edge detection", Proc. Royal Society of London, B, 207, 187–217.
[14] N. Senthilkumaran and R. Rajesh," Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.
[15] M. Stricker and M. Orengo, "Similarity of color images", In SPIE Conference on Storage and Retrieval for Image and Video Databases III, volume 2420, pages 381392, Feb. 1995.
[16] Maheswary Priti, Srivastav Namita &,, Retrieving Similar Image Using Color Moment Feature Detector and K-means Clustering of Remote Sensing Images,, International Conference on Computer and Electrical Engineering, 2008

Author Profile

Annesha Malakar has received her B. Tech degree in Information Technology from West Bengal University of Technology. Presently she is perusing her M.E degree from School of Education Technology, Jadavpur University, Kolkata. Her research interests comprise of Content Based Image retrieval, image Clustering, and Image processing.

Joydeep Mukherjee obtained M. Tech degree from Jadavpur University. Currently he works as an Asst. Prof. in School of Education Technology, Jadavpur University, Kolkata. His research interests comprise of Digital image processing, Character Recognition, Pattern Recognition, E Learning, M- Learning.