Content based Image Retrieval using Clustering

Ashish Kumar Raikwar
M.Tech.
Netaji Subhas Institute of Technology
New Delhi, India

Satbir Jain
Department of Computer Engineering
Netaji Subhas Institute of Technology
New Delhi, India

ABSTRACT

To retrieve appropriate information from large image datasets, Content Based image retrieval (CBIR) is a popular approach. In this paper we use binary clustering simultaneously on target and query images to retrieve color difference. One also measure geometric spreadness of each color, using coordinate information of clusters and used it with color difference with some weighted. Experimental results show that CBIRC gives better result as compare to other binning methods and ACE.

General Terms

CBIR, Clustering, Quantization error.

Keywords

CBIR; binary clustering; geometric spreadness.

1. INTRODUCTION

In the world of technology, information management is become very crucial. As the people know about information technology the use of the computer and digital devices has been increases. People download picture or capture image from digital camera and then upload on the internet. That means we can say that image data are generated very rapidly that create very large image database. This data need to be stored in such a way, so that it can be access easily and quickly.

An image is a group of pixel that represents the object or region in the image. Many color models are exist to represent the image like RGB, HSV, HSL etc. RGB color model are simple model that uses the primary color (red, green, blue) to represent the color intensity value of the pixel in an image. A color image may be described as three layered image of red, green and blue plane.

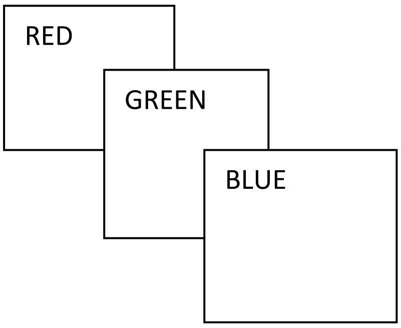

Fig. 1: RGB Color Model

In CBIR system images are searched by providing quarry image as a input. Mostly CBIR System consists of some basic steps that need to be followed. These are image pre-processing, feature extraction and similarity measure or finding distance[1].In the pre-processing stage noise from the image are removed and image converted in some suitable format so that operation on the image can be performed easily and quickly. In the next step various image feature are extracted on the basis of image property like color, texture, shape, edge etc. The number of extracted feature and there type plays an important role on the result and performance of the system. All image features are stored in the feature database. At last query image feature compared to feature database and similar images are retrieved. In the last step distance measure between the target image and quarry image are computed using some distance measure e.g. Euclidian distance, Root-mean distance, City block etc. In our proposed method called Content Based image retrieval using Clustering (CBIRC) we use binary clustering to cluster the color data. In this method quarry image and target image are combined and then difference is calculated. In addition to the color feature in CBIRC, geometric distribution of colors is also taking in account to compare the images.

2. RELATED WORK

Image histogram has been used from many years to retrieve the image and video. Various systems used it for similarity measure e.g. QBIC [2], WebSeek [3] etc. Histograms are popular because they are simple and involve easy computation. Image features are represented by the histogram and each bin represents the number of pixel for a particular color. The distance between images is calculated as the distance between the histogram of image. However, some drawback is associated with the histogram, as they are not containing much information about color distribution and for two image that are completely different may have the same histogram.

Stricker and Orengo [4] suggest a method for color indexing in which three color moments i.e. average, standard deviation and skewness are used to represent the color distribution of an image. Then similarity major between two images is the sum of the difference between these color moments with some weights.

Dominant color descriptor is popular technique to represent the color feature and is used by many researchers for image comparison and retrieval. Same types of color are represented by their dominant color, by this way the performance of the system can be improved. Y. Deng et al. [5] used a color descriptor that contains representative color and its percentage in a region. Image segmentation are apply that form different color regions then with the help of clustering obtained the representative colors for each region.

Local Fourier Transform (LFT) used by Hui Yu et al. [6] to represent texture information in the image. Total eight templates are generated by applying LFT on the image. These templates characterized the image. Then first and second moments are calculated and feature vector are extracted. To overcome the problem in HSV color space ,three dimensional color vector are suggest to used for representing the color.

Man Wong et al. [7] describe a problem in RGB based histogram intersection method in which images that are visually similar but different brightness or intensity value may be considered less similar and vice versa. One another problem is the case where two visually similar color of different value are not matched because their intersection area of their histogram are very less or may be nil. To overcome this problem a Merged Histogram Method (MHM) is proposed where instead of making individual color palette a common palette is constructed by intra and inter palette merging.

A cluster based image retrieval scheme is introduced by Chen, Wang and Krovetz [8] called cluster-based retrieval of images by unsupervised learning (CLUE). It retrieves the image clusters instead of a set of ordered images. Image cluster are formed by using the unsupervised learning that consider image features as well as how the image are similar to each other semantically.

Fuzzy color histogram is proposed by El-Feghi et al. [1] in which membership value for each color pixel value is calculated using membership function. In this method each pixel may be belongs to more than one color set with some membership value. Manhatin distance is used to find the distance measure between images.

Quynh,Tao and Gyang [9] gives a method for similarity major between image is proposed called Histogram Graph (HISG) .Image are divided in the sequence of sectors and calculate histogram for each color where each bin represented by vertex of the graph. Sum of the minimum cost matching is the distance between two images.

In [10] author gives Adaptive Feature Extraction that is based on BQMP [11] technique in which quaternion moments is used to represent the color data. First color cluster are construct by iteratively calling BQMP. A new distance measure called comparing histogram by clustering (CHIC) is introduced to find the difference between histogram.

3. PROPOSED WORK

For effective image retrieval clustering are used in extensively manner by many researcher. In our work we use binary clustering on color data. We resize the image in equal size and format so that color processing would be easy and effectively work on all images. Let I_q is the quarry image i.e. image to be search and I_t is the image from the image database i.e. target image. In the proposed method all three matrix plane of I_q and I_t are combined row by row and considered as a single image matrix *I* as shown in figure (2) and (3). Then arrange these matrix planes in one column index form using (1) such that entries of each column sequentially fill up by rows i.e. first quarry image are placed than after target image are place that we call image *I*. On this image matrix *I* apply clustering. To identify any pixel belongs to I_q or I_t a variable d_p are used that store total number of pixel in I_q or I_t. In matrix plane *I* if the index value of any pixel is greater than d_p then it belong to I_t otherwise it belong to I_q. Then apply binary clustering on the data set *S* so that S_i number of clusters are generated and each S_i contain sub cluster S_{i_q} and S_{i_t} .The sub cluster S_{i_q} and S_{i_t} contains quarry image and target image pixels respectively. After the clustering process we find the difference between the images by absolutely adding the discrepancy of each cluster. This method also save the index value in the index cluster G_i. G_i is generated corresponding to each and every color value of the cluster S_i called index cluster and used for retrieving the geometrical feature of the image. On these cluster we calculate geometric mean and geometric standard deviation and then find the geometric difference between the images. The index clusters are stored corresponding to color cluster of image *I*.
We classify the color data in two class say x_0 and $x_{1.}$ Actually by doing this we are finding out the hyper-plane that will divide the color data set *S* in x_0 and x_1.

As shown in figure 3 quarry image pixel $R_{(x_i,y_i)_q}$, $G_{(x_i,y_i)_q}$ and $B_{(x_i,y_i)_q}$and target image$R_{(x_i,y_i)_t}$, $G_{(x_i,y_i)_t}$ and $B_{(x_i,y_i)_t}$ respectively are combined and treated as a single image *I*.

Coordinate Index Conversion: Any pixel value *(x_i, y_i)* are converted in index form. If *M* is the matrix with *n* rows and *m* columns than index formulas are

$$I = m * (x - 1) + y \qquad (1)$$

$$x = (I - y)/(m + 1) \qquad (2)$$

$$y = \begin{cases} m \bmod I & when\ m \bmod I > 0 \\ m & otherwise \end{cases} \qquad (3)$$

Where *I* is the index value, *x* is the row of the element i.e. *x* coordinate and *y* is the column of element i.e. *y* coordinate.

R_{11}	**R_{12}**	**R_{12}**
R_{21}	**R_{22}**	**R_{23}**
R_{31}	**R_{32}**	**R_{33}**

Red

G_{11}	**G_{12}**	**G_{13}**
G_{21}	**G_{22}**	**G_{23}**
G_{31}	**G_{32}**	**G_{33}**

Green

B_{11}	**B_{12}**	**B_{13}**
B_{21}	**B_{22}**	**B_{21}**
B_{31}	**B_{32}**	**B_{33}**

Blue

Quarry image

R'_{11}	**R'_{11}**	**R'_{11}**
R'_{21}	**R'_{21}**	**R'_{21}**
R'_{31}	**R'_{31}**	**R'_{31}**

Red

G'_{11}	**G'_{11}**	**G'_{11}**
G'_{11}	**G'_{11}**	**G'_{11}**
G'_{11}	**G'_{11}**	**G'_{11}**

Green

B'_{11}	**B'_{11}**	**B'_{11}**
B'_{11}	**B'_{11}**	**B'_{11}**
B'_{11}	**B'_{11}**	**B'_{11}**

Blue

Target Image

Figure: 2

For the input color data set *S* we use variance σ_t as a threshold value. One can decide value of σ_t empirically. For splitting any cluster $S_{i,}$ the variance σ_i of that cluster must be greater than σ_t. To divide the data set *S* of image in two subsets S_k and S_{k+1} assign some index value k_i for each and every pixel p_i of image *I*. These index values are also used to calculate geometric difference.

Initially *S* contains all color values of image *I*. To split image cluster *S* in two classes calculate the mean and variance of the Image *S*. If the variance σ_i is greater than some threshold variance σ_t then it is a splitable cluster thus split this cluster in S_k and S_{k+1}. This process is repeats until we meet all the clusters are unsplitable.

Now calculate mean $\bar{x}_m$ for image *I*.

$$\bar{x}_m = \frac{1}{n}\sum_{i=1}^{n}(R_i, G_{i,}B_i)$$

$$\bar{x}_m = \frac{1}{n}\sum_{i=1}^{n} p_i$$

Than calculate difference D_i between $\bar{x}_m$ and x_i where x_i represents the color value of (R_i, G_i, B_i) for pixel p_i.

$$D_i = (x_i - \bar{x}_m)$$

$$D_i = (R_i - R_m) + (G_i - G_m) + (B_i - B_m)$$

If the D_i is negative than the pixel are placed in sub cluster S_k otherwise it will placed in sub cluster S_{k+1}.Than calculate variance of cluster S_k and S_{k+1} i.e. σ_{s_k} and $\sigma_{s_{k+1}}$ .We first split that cluster whose variance is maximum. Suppose $\sigma_{s_{k+1}}$ is greater than σ_{s_k} and σ_t then split this cluster. Otherwise if $\sigma_{s_{k+1}}$ is not greater than σ_t than S_{k+1} will not split and the splitting process will stop. Similarly if σ_{s_k} is greater than σ_t then split S_k. This procedure repeats for each cluster S_{k+n} until we meet all the unsplitable clusters i.e. the variance of all cluster is less than threshold value σ_t. Each cluster represents the same type colors of the image I. Then find the discrepancy of each cluster.

$$Dis_i = \left|car\left(S_{i_t}\right) - car(S_{i_q})\right|$$

$$Color\ Diff = \sum_{i=1}^{n} Dis_i$$

By adding discrepancy of all clusters we get the color difference between the I_q and I_t.

In any cluster S_i some pixel are from quarry image and some pixel are from target image. Corresponding to each cluster S_i, index cluster G_i is generated that contain index value k_i corresponding to each pixel p_i. The index value k_i is converted in coordinate (x_i, y_i) using equation (2) and (3).

R_{11}	**R_{12}**	**R_{12}**	**G_{11}**	**G_{12}**	**G_{13}**	**B_{11}**	**B_{12}**	**B_{13}**
R_{21}	**R_{22}**	**R_{23}**	**G_{21}**	**G_{22}**	**G_{23}**	**B_{21}**	**B_{22}**	**B_{21}**
R_{31}	**R_{32}**	**R_{33}**	**G_{31}**	**G_{32}**	**G_{33}**	**B_{31}**	**B_{32}**	**B_{33}**
R'_{11}	**R'_{11}**	**R'_{11}**	**G'_{11}**	**G'_{11}**	**G'_{11}**	**B'_{11}**	**B'_{11}**	**B'_{11}**
R'_{21}	**R'_{21}**	**R'_{21}**	**G'_{11}**	**G'_{11}**	**G'_{11}**	**B'_{11}**	**B'_{11}**	**B'_{11}**
R'_{31}	**R'_{31}**	**R'_{31}**	**G'_{11}**	**G'_{11}**	**G'_{11}**	**B'_{11}**	**B'_{11}**	**B'_{11}**

Red Green Blue

Figure: 3

For x and y co-ordinate of image I_t, $\bar{x}_{gm_{t_i}}$ and $\bar{y}_{gm_{t_i}}$ are the geometric mean respectively.

$$\bar{x}_{gm_{t_i}} = \sqrt[car\,(GS_{t_i})]{\prod_{\forall x_{t_i} \in GS_{t_i}} x_{t_l}}$$

$$\bar{y}_{gm_{t_i}} = \sqrt[car\,(GS_{t_i})]{\prod_{\forall y_{t_i} \in GS_{t_i}} x_{t_m}}$$

Similarly for image I_q geometric mean $\bar{x}_{gm_{q_i}}$ and $\bar{y}_{gm_{q_i}}$ will be

$$\bar{x}_{gm_{q_i}} = \sqrt[car\,(GS_{q_i})]{\prod_{\forall x_{q_i} \in GS_{q_i}} x_{q_l}}$$

$$\bar{y}_{gm_{q_i}} = \sqrt[car\,(GS_{q_i})]{\prod_{\forall y_{q_i} \in GS_{q_i}} x_{q_m}}$$

Where $car(GS_{t_i})$ and $car(GS_{q_i})$ are the cardinality of index sub cluster GS_{t_i} and GS_{q_i} respectively.

For each and every unsplitable cluster calculate the geometric standard deviation (GSTD) with respect to geometric mean of quarry image pixel cluster and target image pixel cluster that are considering as a sub clusters of the cluster S_i. Geometric Standard Deviation $GSTD_{t_i}$ for I_t will be

$$GSTD_{t_i} = \sqrt{\sum_{\forall x_{t_l} \in GS_{t_i}} \left(x_{t_l} - \bar{x}_{gm_{t_i}}\right)^2 + \sum_{\forall y_{t_m} \in GS_{t_i}} \left(y_{t_m} - \bar{y}_{gm_{t_i}}\right)^2}$$

And $GSTD_{q_i}$ for I_q

$$GSTD_{q_i} = \sqrt{\sum_{\forall x_{q_l} \in GS_{q_i}} \left(x_{q_l} - \bar{x}_{gm_{q_i}}\right)^2 + \sum_{\forall y_{q_m} \in GS_{q_i}} \left(y_{q_m} - \bar{y}_{gm_{q_i}}\right)^2}$$

Than calculate the difference between geometric standard deviation of the sub cluster. Do this for all unsplitable clusters and by absolutely adding their difference that show the geometrical color discrepancy.

$$GD_i = \left\{1 - \frac{Dis(S_i)}{car(S_i)}\right\} * \left\{abs\left(GSTD_{t_i} - GSTD_{q_i}\right)\right\}$$

$$GD_{diff} = \sum_i GD_i$$

Adding this Geometrical color discrepancy and color difference of the image will be the total difference between the quarry image and target image.

$$Total\ \ Diff = Color\ diffrence + \ \alpha.Gm$$

Where $G_m$the Geometrical difference and value of is α is selected between zero and one.

In this method geographical information of pixel are also used to classify images more effectively and results are more accurate.

4. EXPERIMENTS

Our experimental result performed on Intel Core2Duo processor of 2.10 GHz with 2 GB RAM in windows XP Environment on MATLAB 7.6.0.324 .We divide all the image in six category . Each and every image belongs to one of the class. We compare our results with the ACE method in the terms of precision graph and figure (6) shows that CBIRC gives the better result.

The performance of color retrieval method measured in terms of average retrieval precision i.e. each time a quarry image select from the database and after processing retrieved k matched image from the database. The average retrieval precision can be defined as

$$\eta = \frac{\sum_{i=1}^{j} ni}{\sum_{i=1}^{j} k}$$

Where j is the no images in the database and n_i is the number of returned image falling into correct image class of quarry image *i*.

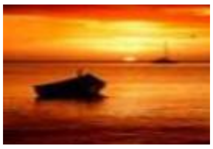

Figure 4: Input image

\

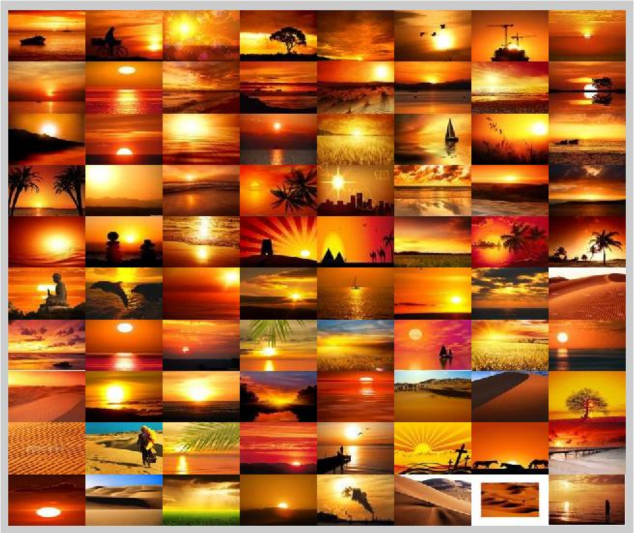

Figure 5: Output of CBIRC

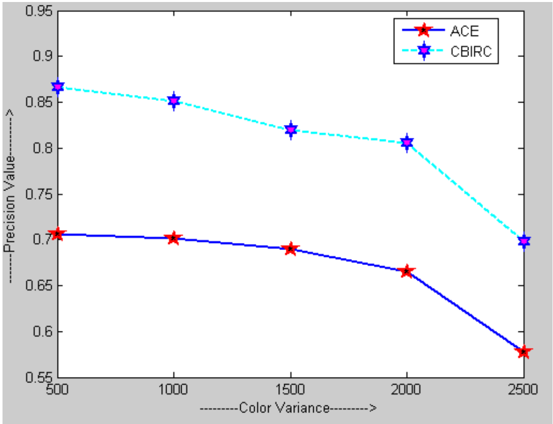

Figure 6: Precision Comparison between CBIRC and ACE (VC)

5. CONCLUSION

For Content Based Image retrieval various image feature are used to compare the image. These features may vary as per requirements. More the feature are selected there are chances that results are closer and accurate but it may lead to increase complexity.

CBIRC apply binary clustering simultaneously on target and query images. Binary clustering generates color and their corresponding index cluster. On the basis of these clusters we calculate color and spreadness differences of corresponding colors. Lastly we add these two differences with appropriate weighted i.e. final difference between images. Experimental result shows that average performance of CBIRC are 14% better than ACE.

REFERENCES

[1] I. El-Feghi, H. Aboasha, M. A. Sid-Ahmed, M. Ahmadi "Content-Based Image Retrieval Based on Efficient Fuzzy Color Signature" 2007 IEEE Transaction.

[2] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, G. Taubin "The QBIC Project: Querying Images By Content Using Color, Texture, and Shape" SPIE Vol. 1908 (1993) pp : 173-187.

[3] John R. Smith, Shih-Fu Chang "Visually Searching the Web for Content" July-September 1997 IEEE Multimedia: pp: 12-20.

[4] M.Stricker and M.Orengo, "Similarity of color images", Storage and Retrieval for Image and Video Databases III (SPIE) 1995: pp: 381-392.

[5] Yining Deng, Manjunath, B.S., Kenney, C., Moore, M.S., Shin, H.,"An efficient color representation for image retrieval," IEEE Transactions on Image Processing, vol.10(1), pp:140 – 147, Jan. 2001

[6] Hui Yu; Mingjing Li; Hong-Jiang Zhang; Jufu Feng," Color texture moments for content-based image retrieval," Proc. ICIP 2002, vol. 3 , pp:929-932, 24-28 June 2002.

[7] Ka-Man Wong, Chun-Ho Chey, Tak-Shing Liu, Lai-Man Po," Dominant color image retrieval using merged histogram," in Proc. ISCAS '03, vol. 2, pp: 908- 911,25-28 May 2003.

[8] Yixin Chen, James Z. Wang, Robert Krovetz," Content-Based Image Retrieval by Clustering," MIR'03, November 7, 2003, Berkeley, California, USA, ACM.

[9] Nguyen Huu Quynh, Ngo Quoc Tao, Ngo Truong Giang, "An efficient method for content based image retrieval using histogram graph", IEEE 10th Intl. Conf. On Control, Automation, Robotics and Vision, pp: 875-876, 17-20 December 2008.

[10] Wei-Ta Chen, Wei-Chuan Liu, and Ming-Syan Chen, Adaptive Color Feature Extraction Based on Image Color Distributions,IEEE Transactions on image processing, pp:2005-2016 vol. 19, no. 8, august 2010.

[11] Soo-Chang Pei, Ching-Min Cheng "Color Image Processing by Using Binary Quaternion-Moment-Preserving Thresholding Technique" IEEE transactions on image processing, pp:616-628 vol. 8, no. 5, may 1999.