2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)

Content Based Image Retrieval using Statistical Features of Color Histogram

Naushad Varish
Department of Computer Science and Engineering
Indian School of Mines
Dhanbad-826004, India
E-mail: naushad.cs88@gmail.com

Arup Kumar Pal
Department of Computer Science and Engineering
Indian School of Mines
Dhanbad-826004, India
E-mail: pal.ak.cse@ismdhanbad.ac.in

***Abstract*—Content based image retrieval (CBIR) is the process of searching similar images from an image database based on the visual contents of the input query image. In CBIR, color is considered as one of the prominent features of the image data, so in this paper, the authors have presented a CBIR technique using color based feature. Since a color image, consists of three basic color components, i.e. red, green and blue, so in this work, we have given the same importance on all three color components during image retrieving process. In the presented CBIR technique, initially we have constructed three probability histograms for each color component and subsequently the histograms are divided into several numbers of significant bins and from each bin, we have computed several statistical values like standard deviation, skewness and kurtosis. The computed statistical values are used as extracted features of the image data. The processing cost of the presented CBIR technique is significantly low. The technique has been tested on standard image databases and satisfactory results have been achieved.**

Keywords—Content based image retrieval; Histogram; Precision-Recall; Statistical parameters;

I. INTRODUCTION

The multimedia based applications have become popular due to the rapid advancement of Internet technology and the digital devices. As a result, the volume of digital image libraries obtained through different type of sources like social networking sites, multimedia cameras, multimedia mobiles, internet etc. has grown tremendously. So there is a demand for suitable searching techniques to retrieve meaningful image data from that large volume of digital image libraries. Image retrieval techniques are used to maintain and retrieve the images in the database. In general, two approaches are used to maintain and retrieve the image data from any image database.

A traditional keyword or text based matching approach for retrieval of images from a large image database, is called a textual method which is based on manual annotation of images with descriptive keywords. So these traditional image retrieval systems used the traditional database management method to index and retrieval of images from a large database based on some simple attributes such as the image num1ber and text description. Traditional methods of image indexing have proven to be insufficient, laborious, and extremely time consuming. The queries are limited in such type of systems and sometime query results are not perfect because features of images are not described completely, effectively and accurately. Hence, to overcome these types of limitations, new techniques have been developed called Content based image retrieval (CBIR) [1-5] techniques for indexing and retrieving images from database.

CBIR performs retrieval process based on the visual contents of the image data. The straightforward image to image searching mechanism is not considered in CBIR. Such approach is not the practically feasible to implement it in any real time applications because image data are comparatively huge in size. So in CBIR, we require suitable feature extraction techniques from the image data so that meaningful and relevant images can be retrieved based on those extracted image features. A number of CBIR techniques were developed based on considering the significant feature like color [6], texture [7] and shape [8-10]. Color feature is widely used in CBIR techniques since it is one of the most prominent low level features and it is also invariant to rotation, scaling, and other spatial transformations on the images. In general the histogram matching based CBIR techniques is relatively simple and faster. Swain et al. [11] method was based on color histogram and similarity measure was done using histogram intersection distance metric between the histograms of images. Another histogram based CBIR was proposed by F. Malik et al.[12] where, they converted the color image into grayscale and the obtained grayscale was further preprocessed by Laplacian filter. The histogram of the processed image was used in CBIR. Wang et al. [13] proposed a method for image retrieval by concatenation of color features like most dominant colors of region, texture features and shape features that are based on pseudo Zernike moments. G. H. Liu et al. [14] used color difference histogram in CBIR. H. B. Kekre et al. [15] also devised CBIR techniques using color histograms and some statistical parameters. Murala et al.[16] proposed CBIR is based on the features like color histogram and Gabor wavelet transform. In [17], the authors have decomposed image into 16 non overlapping color blocks and from each

978-1-4673-6823-0/15/$31.00 ©2015 IEEE

color blocks, they have computed various statistical values as a component of a feature vector. In [18], the authors have suggested a CBIR based on color moment and Gabor texture feature. M. Singha et al. [19] proposed two methods where in first method purely based on color histogram in spatial domain while the second method is based on wavelet domain. A histogram was constructed from the color image after color quantization process. In CBIR, the feature vector was computed from the histogram of the color image. In this paper, we have presented CBIR techniques using color histogram. The straight forward histogram comparison is not suitable in any real time scenario. So we have computed several statistical values from the histogram. The retrieving was done based on the extracted statistical values. The scheme is suitable and easy to implement in any real time applications.

The rest of the paper is organized as follows: in section 2, we have discussed in briefly the working principle of the content based image retrieval. In section 3, we have presented the idea of some statistical parameters. The presented scheme and the experimental results are depicted in section 4 and 5 respectively. Finally, section 6 concludes the paper.

II. Content Based Image Retrieval

In CBIR, indexing and retrieval of images is based on low level visual features of images such as color, texture and shape. Basically CBIR is performed in two steps: indexing and searching. In indexing, the low level features are extracted from image database and stored in the form of feature vector to create a new database which is known as feature database. In searching, initially the same feature extraction algorithm is employed in the query image to extract the feature vector and subsequently the similarity is measured with the feature database using some distance measure like Euclidean distance. The best result is obtained those having minimum similarity distance. The schematic diagram of the Content-based image retrieval (CBIR) techniques is shown in Figure 1.

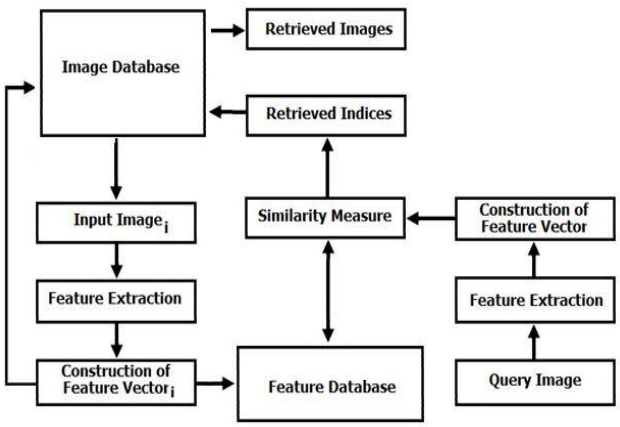


Figure1. Content-based image retrieval System

III. Statistical Parameters

Performance of any CBIR techniques depends on the extraction of suitable features of the images. CBIR will give efficient and effective result if we consider more features of the images, but as a result overall computational cost will be high. So from a practical point of view, the dimension of the feature vector of the image should be selected appropriately so that the performance may not be affected impressively. Therefore, in CBIR, selection of the suitable dimension of the feature vector is challenging task. In this paper some statistical features of the images have been measured. In some the cases, the statistical features have produced significant results in the classification of the datasets. Some commonly used statistical parameters [20] are mean, standard deviation, skewness and kurtosis etc. These statistical parameters can be computed directly from an image histogram. Initially, the original histogram is converted into a form of a normalized histogram where X-axis denotes the intensity level, r_i and Y-axis denotes the estimated probability $p(r_i)$ of level r_i. The mean (μ) for the range of intensity value $[LB,UB]$ is computed as

$$\mu = \sum_{i=LB}^{UB} r_i p(r_i) \tag{1}$$

The standard deviations (σ), skewness (γ), and kurtosis (κ) for the range of intensity value $[LB,UB]$ are obtained as follows:

$$\sigma = \sqrt{\sum_{i=LB}^{UB} (r_i - \mu)^2 p(r_i)} \tag{2}$$

$$\gamma = \frac{1}{\sigma^3} \sum_{i=LB}^{UB} (r_i - \mu)^3 p(r_i) \tag{3}$$

$$\kappa = \frac{1}{\sigma^4} \sum_{i=UB}^{UB} (r_i - \mu)^4 p(r_i) \tag{4}$$

The standard deviation shows the contrast of the image in the particular significant bin of the histogram of the image. It is used to measure the distribution of the intensity values about mean in each bin of the histogram. If the value of the standard deviation in a particular block of the image is less the value of the other block in the image, it represents that high contrast in that particular block of the image than the other. The skewness is the measure of the skewed intensity values of the image in each block of the image about mean of that block. If the value of the skewness is negative in the particular region of the image, then it represents that most of intensity values lie on the right side of the mean than the left side and tail of the intensity values longer and lie on the left side of the mean of that region of the image. If the value is positive, then most of the intensity values lie in the left side of the mean in a region and tail of the intensity values more skewed in the right side of the mean. If skewness is zero then it represents that distribution of the intensity values about mean is equal. The kurtosis is used to calculate the peak of the distribution of the intensity values of the image about mean in a significant bin of the histogram. If Kurtosis has high value, then it represents the sharp peak distribution of the intensity values about mean and has longer and fat tail and kurtosis with low value represents flat distribution of intensity values with short and thin tail.

Sometime statistical parameter like mean does not contribute appropriate results in similarity measurement among dataset. Figure 2 shows two types of image blocks where D1 image blocks have smooth region but in D2 image block a horizontal edge is prominent. However, both the image blocks have same mean value, i.e. 38.89 but standard deviations, skewness and kurtosis of dataset for D1 are 6.5085, 0.4289 and 1.9741 and for D2 are 31.1024, 0.6546 and 1.4947 respectively. Therefore, these parameters are capable to find out the variation between D1 and D2 image blocks. So in this paper, we have considered statistical parameters as standard deviations, skewness and kurtosis in the presented CBIR technique.

D1		
30	35	35
35	45	40
45	50	35

D2		
15	25	15
80	80	80
15	25	15

Figure 2. Image Block

IV. Statistical Feature Extraction based CBIR Technique

The performance of CBIR specially depends on developing of suitable and fast image feature extraction technique. So feature extraction in CBIR technique is considered as a fundamental step. The schematic diagram of the employed feature extraction technique for CBIR is shown in Figure 3. The algorithmic steps of the feature extraction process from an image are presented as follows:

Algorithm 1: Feature Extraction

Begin

Step 1. Take RGB color image as an input and decompose into its three color components, i.e. Red(R), Green (G) and Blue B) respectively where each color component has L intensity levels.

Step 2. Construct the histograms H_R, H_G and H_B from the color components R, G and B respectively.

Step 3. For each color component, compute the probability histogram as follows

$$P(r_i) = \frac{Number\ of\ Pixels\ in\ r_i}{(Weidth \times Height)} \qquad \forall i \in [0, L-1]$$

Where $p(r_i)$ represents the relative frequency or probability of r_i -th intensity value and the range of intensity value is $[0, L-1]$.

Step 4. Divide the probability histogram into $n(<< L)$ number of non-uniform bins where the i-th bin is computed by

$$Bin_i = \sum_{j=LB}^{UB} p(r_j) \le \frac{1}{n} \qquad \forall i \in [0, n-1]$$

This implied that the each modified bins is obtained by taking the cumulative probability of $p(r_i)$ in such way that the probability will be less than $1/n$.

Step 5. Compute for each color component, the standard deviation, skewness and kurtosis from $Bin_{iC} \qquad \forall i \in [0, n-1]$ where 'C' represents R, G and B color components respectively.

Step 6. The feature vectors obtained from Step 5 is represented as

$$fv = \{\sigma_{jR}, \sigma_{jG}, \sigma_{jB}, \gamma_{jR}, \gamma_{jG}, \gamma_{jB}, \kappa_{jR}, \kappa_{jG}, \kappa_{jB}\}, \ \forall j = 0,1,...n-1$$

End

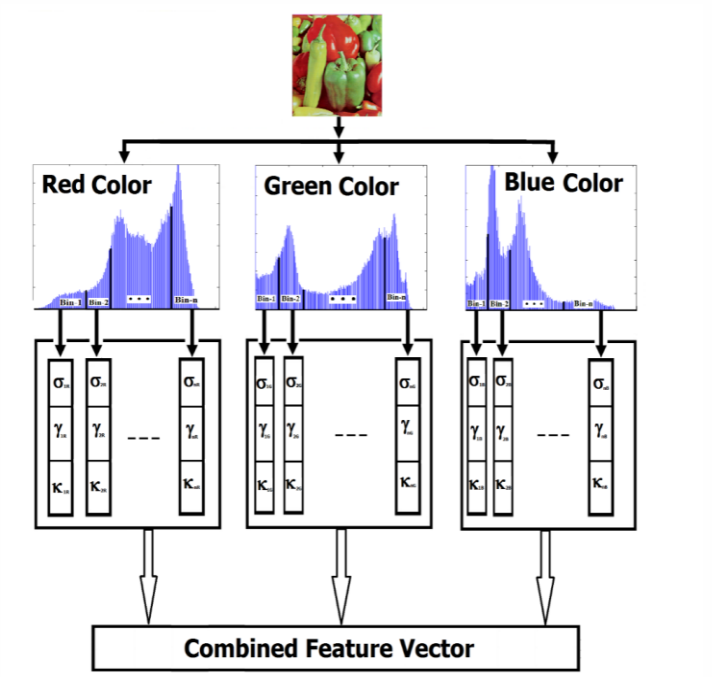


Figure 3.Schematic diagram for feature extraction process

The CBIR is developed in two major steps i.e. proper training of the images and perform searching or retrieving of relevant images based on the trained image data. Feature extraction is the part of training process where same feature extraction algorithm is deployed on both the query image and the images of Database. In searching, suitable distance measurements are considered. The major algorithmic steps of the presented CBIR are as follows:

Algorithm 2: Proposed CBIR Model

Begin

Step 1: Construct feature vectors for the query image and the database images using the feature extraction technique (using Algorithm 1).

Step 2: Compute the distance between feature vector of the query image and the feature vectors of the database images for the similarity measurement.

Step 3: Sort the distance in non-decreasing order and select top N images having minimum distances.

End

V. Experimental Results and Performance Measurement

The presented method is tested on the Corel image database [21] which is freely available in the internet for the researchers. The database contains 1000 images of 10 categories. Each category has 100 images of the type people images, horse images, mountain images, elephant images, building images, bus images, rose images, food images, beach images and dinosaur images. In the presented method, the images are taken from the database one after another for extracting the features of the images to form feature vectors and stored these feature vectors in the feature database. Suitable distance measurement is used for similarity matching between the query image and database images on the basis of the computed feature vectors. In this paper we have considered the Euclidean distance as similarity measurements and is defined as

$$\Delta D = \sum_{z=1}^{s} \sqrt{(F_z(Q) - F_z(P))^2}, \ \forall z = 1,2...s \qquad (5)$$

Where $F(Q)$ and $F(P)$ are the feature vectors of the query image and database image respectively and s is the dimension of the feature vectors. The experimental result is also assessed with others distances like City block distance and Chebychev distance. The City block distance is defined as mathematically

$$\Delta D = \sum_{z=1}^{S} \left| F_z(Q) - F_z(P) \right| \ \ \forall z = 1,2...s \qquad (6)$$

And the Chebychev distance is computed as follow

$$\Delta D = Max\left\{ \left| F_1(Q) - F_1(P) \right|, \left| F_2(Q) - F_2(P) \right|, ..., \left| F_z(Q) - F_z(P) \right| \right\} \qquad (7)$$

The performance of the CBIR can be measured by two metrics. One is called precision and other is called Recall. As precision increases better result is obtained and decreases bad result is obtained. Recall is just opposite of precision as a result high recall gives bad result and low recall gives better result. These two accuracy measurement metrics are defined given below. Precision (P) and Recall(R) can be defined by the mathematically

$$P = X / Y \qquad (8)$$

$$R = X / Z \qquad (9)$$

where 'X' is the total relevant images retrieved, 'Y' is the total number of images retrieved and 'Z' is the total relevant images available in the database.

In the experiment, the histogram of each color component is divided into 10 bins and from each bin, we have computed 3 statistical values. As a result, the feature vector length is 90 for three components. We have considered first 10 relevant images those having minimum Euclidean distance. The results in term of precision and recall obtained from experiment for different categories are depicted in Table 1. The result shows the best performance in terms of retrieval accuracy. It is clear from the given table I that the proposed method gives 100% precision values for the categories of horses and dinosaurs images. It gives poor performance for the categories of building and mountain images since it contains various background features. We have also compared the results with some other color feature based CBIR techniques [17-19] where they have considered same image database. Table II shows the relative performances. The average precision and recall values are given for different distance measurement and it shows that Euclidean distance measurements gives better results compare to the others. The average precision for three different distances is shown in Figure 4. The simulation results for the different queries of the individual categories of the images have been performed. The simulation results for the dinosaurs, flowers, buildings and peoples shown in the Figure 5.

TABLE I. Precision and Recall of Proposed Method for Top Ten Retrived Images

SI No.	Category	Precision	Recall
1	Peoples	100	10
2	Beaches	90	9
3	Buildings	50	5
4	Buses	80	8
5	Dinosaurs	100	10
6	Flowers	90	9
7	Horses	100	10
8	Foods	100	10
9	Mountains	50	5
10	Elephants	80	8

TABLE II. Comparision with other Related Techniques

SI No.	Category	M. Imran et al. [17]	CMW [18]	CMR [18]	CH [19]	Proposed
1	Peoples	85	75	75	72	100
2	Beaches	62	46	38	53	90
3	Buildings	54	25	35	61	50
4	Buses	45	67	78	93	80
5	Dinosaurs	100	74	83	95	100
6	Flowers	44	42	61	66	90
7	Horses	67	55	70	89	100
8	Foods	28	67	62	82	100
9	Mountains	21	43	43	47	50
10	Elephants	57	60	45	84	80
	Average	**56.0**	**55.4**	**59.0**	**74.2**	**84.0**

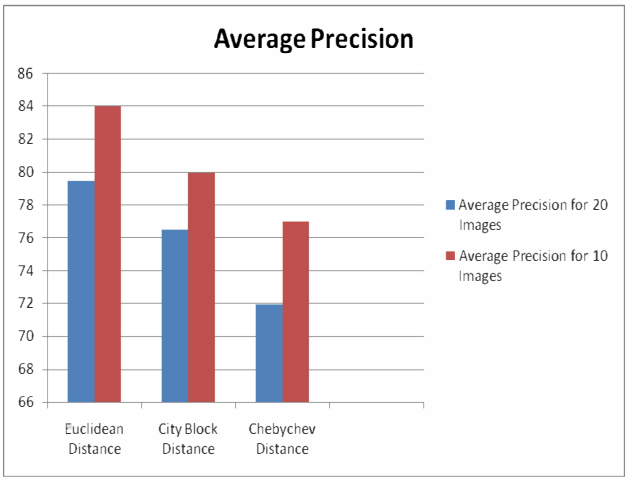
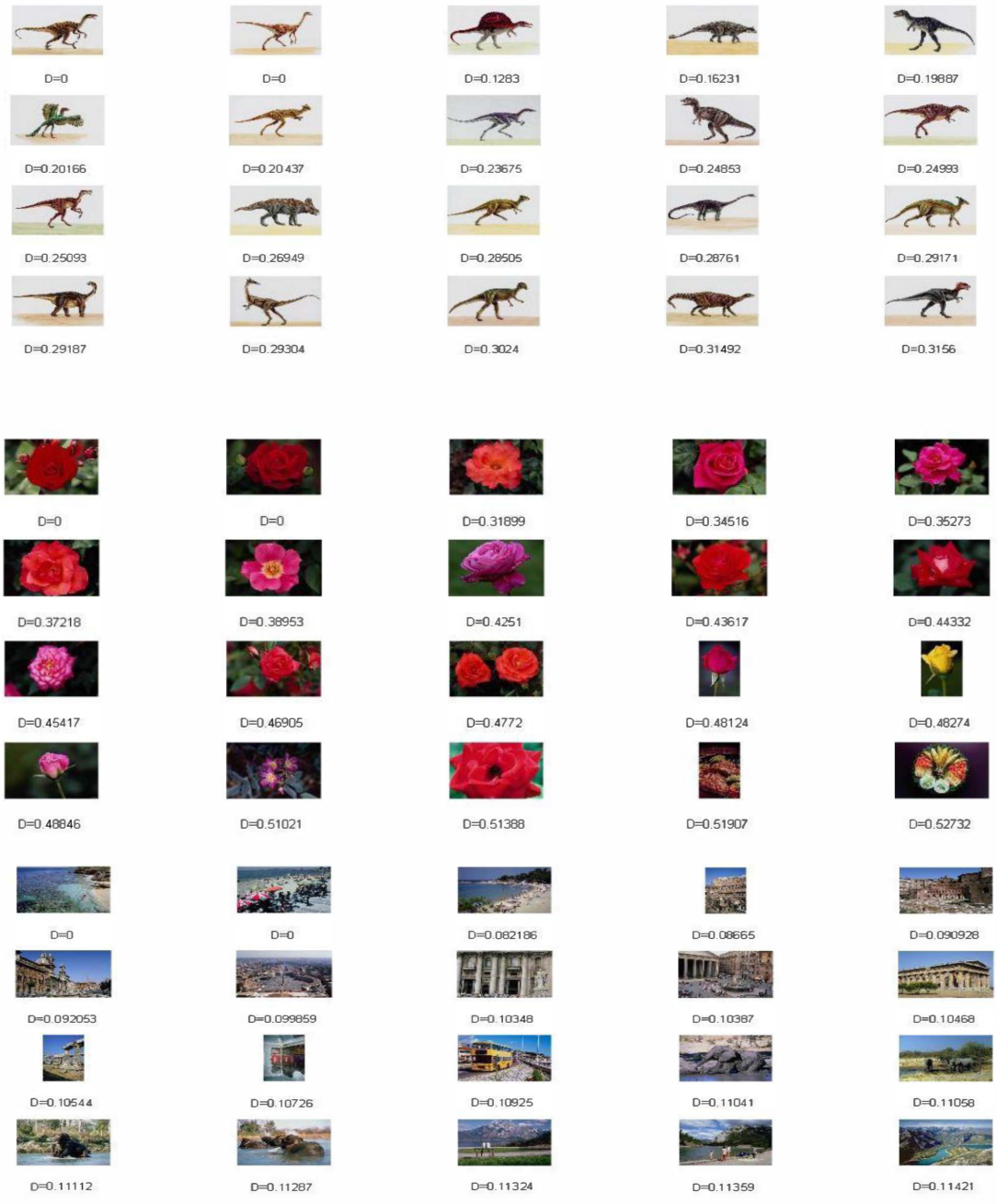


Figure 4.: Average Precision and Recall graph for different distance measurements

VI. CONCLUSION

The proposed method is simple and effective during the retrieval process. It is based on the color histogram which is partitioned into several numbers of non-uniform bins that contain most significant information of the image. Features can be extracted from each significant bins of the histogram for retrieving of the relevant color images from the database. The statistical parameters are calculated directly from the image histogram as a result, it reduces the processing cost. The presented method gives the 100% precision value for horse and dinosaur categories images where for Building and mountain images gives the worst result due to the presence of more details features. However, the performance of this method for other images is acceptable and it is suitable for any real time applications.



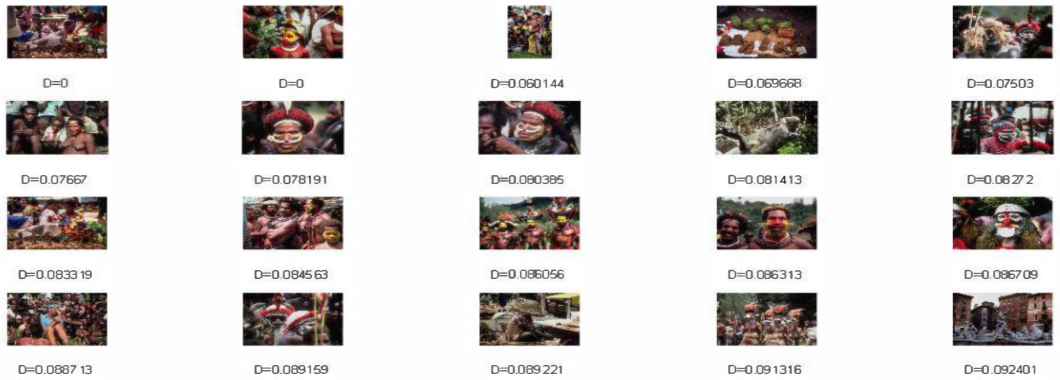


Figure 5.Simulation Results

REFERENCES

[1] M. Kokare, B. N. Chatterji, and P. K. Biswas. "A survey on current content based image retrieval methods." IETE Journal of Research, vol. 48 no. 3-4, pp. 261-271, 2002.

[2] L. Ying, D. Zhang, Lu. Guojun, Ma. Wei-Ying. "A survey of content-based image retrieval with high-level semantics," Pattern Recognition,vol. 40, no. 1, pp. 262–282, 2007.

[3] A. W. M. Smeulders, M. Worring, S. Santini and A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," Pattern Analysis and Machine Intelligence, IEEE Transactions , vol. 22, no. 12, pp. 1349–1380, 2000

[4] A. Vailaya, A. T. Figueiredo, A. K. Jain, and H. J. Zhang, "Image classification for content-based indexing," Image Processing, IEEE Transactions, vol. 10, no. 1, pp. 117-130, 2001.

[5] Tong S. and Chang E., "Support vector machine active learning for image retrieval," In Proceedings of the ACM International Conference on Multimedia, Ottawa, Canada, pp. 107–118, 2001.

[6] N. Shrivastava and V. Tyagi, "An efficient technique for retrieval of color images in large databases," Computers and Electrical Engineering, 2014.

[7] S. D. cheng, X. U. Lan, and L. Y. Han, "Image retrieval using both color and texture features," The Journal of China universities of posts and telecommunications, vol. 14, pp. 94-99, 2007.

[8] R. Krishnamoorthy and S. S. Devi, "Image retrieval using edge based shape similarity with multiresolution enhanced orthogonal polynomials model," Digital Signal Processing, vol.23, no. 2, pp. 555–568, 2013.

[9] N. Alajlana, I.E., Rube, M.S. Kamel and G. Freeman, "Shape retrieval using triangle-area representation and dynamic space warping," Pattern Recognition, vol. 40, no. 7, pp. 1911–1920, 2007.

[10] M. A. Z. Chahooki and N.M. Charkari, "Shape retrieval based on manifold learning by fusion of dissimilarity measures," Image Processing, IET, vol. 6, no. 4, pp. 327-336, 2012.

[11] M. J. Swain and D.H. Ballard, "Color indexing," International Journal of Computer Vision, vol. 7, no. 1, pp. 11-32, 1991.

[12] Malik F., and Baharum B., "Feature analysis of quantized histogram color features for content-based image retrieval based on laplacian filter," International Conference on System Engineering and Modeling, Singapore, 2012.

[13] X. Y. Wang, Y. J. Yong, and H. Y. Yang, "An effective image retrieval scheme using color, texture and shape features," Computer Standards & Interfaces, vol. 33, no. 1, pp. 59-68, 2011.

[14] G. H. Liu, and J. Y. Yang, "Content-based image retrieval using color difference histogram," Pattern Recognition, vol. 46, no. 1, pp. 188-198, 2013.

[15] Kekre H. B. and Sonawane, K., "Use of equalized histogram CG on statistical parameters in bins approach for CBIR," In Advances in Technology and Engineering (ICATE), International Conference, IEEE, pp. 1-6, 2013

[16] Murala S., Gonde A. B. and Maheshwari, R. P., "Color and texture features for image indexing and retrieval," In Advance Computing Conference(IACC), IEEE , pp. 1411-1416, 2009.

[17] M. Imran, R. Hashim, and N. E. A. Khalid, "Color histogram and first order statistics for content based image retrieval," In Recent Advances on Soft Computing and Data Mining, Springer International Publishing, pp. 153-162, 2014.

[18] S. M. Singh and K. Hemachandran, "Content-based image retrieval using color moment and gabor texture feature," International Journal of Computer Science Issues(IJCSI), vol. 9, no. 1, pp. 1694-0814, 2012.

[19] Manimala S. and Hemachandran, K., "Content based image retrieval using color and texture," Signal & Image Processing, an International Journal (SIPIJ), vol. 3, no. 1, pp. 39-57, 2012.

[20] R. C. Gonzalez, R. E. Woods and S.L. Eddins, "Digital Image Processing," 2nd Edition Pearson Prentice Hall Book, 2004.

[21] Li. Jia, and J. Z. Wang, "Automatic Linguistic Indexing of pictures by a statistical modeling approach," Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 25, no. 9, pp. 1075-1088, 2003.