International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Content Based Image Retrieval and Classification Using Principal Component Analysis

Roshani Mandavi[1], Kapil Kumar Nagwanshi[2]

[1]M.Tech Scholar, Computer Science & Engineering, RCET, Bhilai

[2]Associate Professor, RCET, Bhilai

Abstract: ***Content based image retrieval (CBIR) systems has become more popular nowadays in many applications. The main issue in content-based image retrieval system (CBIR) is to extract the image features that represent the image contents in a database. Such an extraction requires a detailed analysis of retrieval performance of image features. In this research we will propose a retrieval system which uses color, texture and shape features of an image that represent the contents of image. Proposed system extract the color, texture and shape feature using color moment, grey-level co-occurrence matrix (GLCM) and Fourier descriptors respectively. After the features are selected, a PCA based classifier classifies an image based on trained feature database and results of this system retrieves the relevant images. It will propose an efficient retrieval system which is able to select the most relevant features to analyze new encountered images thereby improving the retrieval efficiency and accuracy.***

Keywords: CBIR (content based image retrieval), Feature extraction, Color moment, Grey level co-occurrence matrix(GLCM), Fourier descriptor, PCA (principal component analysis), Feature similarity.

1. Introduction

Nowadays the application of internet and www is increasing exponentially and the collection of image accessible by the users is also growing in numbers. During the last decade there has been a rapid increase in volume of image and video collections. A huge amount of information is available, and daily gigabytes of new visual information is generated, stored, and transmitted. However, it is difficult to access this visual information unless it is organized in a way that allows efficient browsing, searching, and retrieval. Traditional methods of indexing images in databases rely on a number of descriptive keywords, associated with each image. However, this manual annotation approach is subjective and recently, due to the rapidly growing database sizes, it is becoming outdated. To overcome these difficulties in the early 1990s, Content-Based Image Retrieval (CBIR) emerged as a promising means for describing and retrieving images. According to its objective, instead of being manually annotated by text-based keywords, images are indexed by their visual contents such as color, shape, texture, and spatial layout.

The importance of content-based image retrieval for many applications, ranging from art galleries and museum archives to picture collections [1], criminal investigation, medical[2][3] and geographic databases, makes the visual information retrieval one of the fastest growing research fields in information technology. Therefore, many content-based retrieval applications have been created for both research and commercial purposes.

CBIR is an application of computer vision techniques to the image retrieval problem that is the problem of searching for digital images in large databases. The task of CBIR systems is to search and browse an image from large database. In CBIR systems the word "Content-based" means, the search searches the contents of image rather than metadata such as tags, keywords and descriptions related with the image. In this topic the term "content" might refer to textures, colors, shapes or any additional information that can be extracted from the image itself. Some CBIR systems have been develop such as QBIC [4][5], MUVIS [6], VisualSEEK [7], SQUID and Photobook [8] in the field of image retrieval research.

The PCA algorithm is used for classification.Various researches have been done on PCA algorithm that providesmore accurate image classifier system than other techniques [9].The main work of PCA is to extracts principal features of an image. These principal features are integrated in predefined class or a single module [10]. Various researchers analyses that the PCA based technique provide better classification and accurate output in the field of computer vision like weather forecasting [9], face identification [11],face recognition [12], feature based image classification [9], medical diagnostics [13], remote sensing images [14], data mining.

The main purpose of this research is to implement the CBIR systems methodology based on contents or low-level features of an image. The property of CBIR systems maintains the large database. In this research the CBIR system trained by the collection of database and the most relevant images are retrieve according to the feature comparison between the feature databases and feature of query image. The data is taken form WANG'S database [15]. Features are extracted using color moment model, GLCM and Fourier descriptor method for color, texture and shape respectively. After feature extraction the PCA calculates principal components from features of both query and trained images and then classifies the query image to its respective class.

2. Methodology

The CBIR system retrieves the image from digital image database on the basis of color or texture or shape. Among all these three features combination of color, texture and shape

feature works very effectively in most situations. According to the figure 1 when a query image is submitted for retrieval purpose, color texture and shape features of the image are extracted and classification with matching operation is performed between the query image feature and image database features, the results for requested query image is then retrieved from the database which is more closes to the query image.

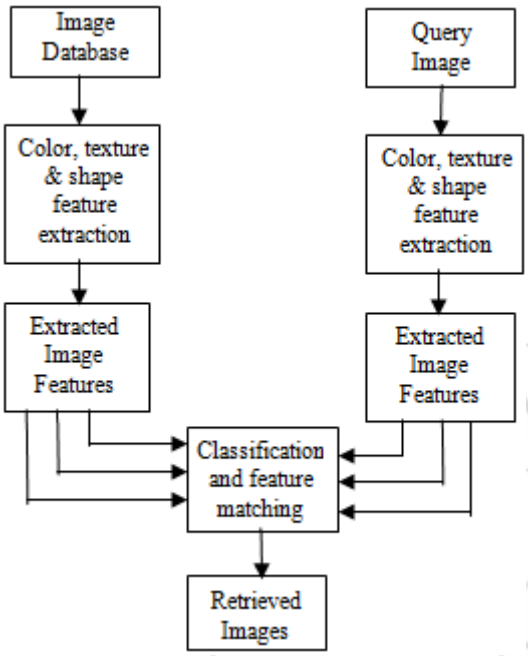


Figure 1: Prototype of CBIR System

In this system the trained database contains 150 images including butterfly, fruit, texture, car and flower with different illuminations taken from Wang's image database and 30 images for testing. The first step of our system is to pre-process the query image. After pre-processing color feature, texture feature and shape feature will be extracted using color moment model in HSV color space, GLCM matrix, Fourier descriptor transformation. After feature extraction the PCA calculates principal components from features of both query and trained images and then classifies the query image to its respective class. The classification method PCA makes our retrieval system to more effective and robust.

A. Color feature extraction using color moment

The main aspects of color feature extraction are the selection of a color space. A color space is a conceptual tool that describes the color capabilities of digital image file. In this research we select HSV color space to extract color feature of an image. First we plot the RGB image to HSV color space. HSV provides the perception representation according with human visual feature. The HSV model, defines a color space in terms of three constituent components first is Hue, the color type vary from 0 to 360 relative to the color red and red primary starting at 0°, passing through the green primary color at 120° and also the blue primary color at 240°, and then back to red color at 360°. Second component is Saturation which represents the "vibrancy" of the color varies from 0 to 100 % and also called the "purity". And last third components are Value which represents the brightness of the color varies from 0 to 100%. We can transform pixel values from RGB color space to HSV color space using equation (1),(2),(3).

$$H = \cos^{-1}\frac{\frac{1}{2}[(R-G)+(R-G)]}{\sqrt{(R-G)^2+(R-B)(G-B)}} \quad \ldots (1)$$

$$S = 1 - \frac{3[\min(R,G,B)]}{R+G+B} \quad \ldots (2)$$

$$V = \left(\frac{R+G+B}{3}\right) \quad \ldots (3)$$

After converting RGB color space to HSV color space, we calculate three central moments of color distribution of a query image in HSV color space. The moments are mean, standard deviation and skewness. A color of any input image can be defined by three or more than three values. Here we convert RGB color image to HSV color image so that moments are calculated for each of these channels for an image. Hence an input image is characterized by 9 moments. We get three moments for each 3 color channels (H, S, and V). The Mean, standard deviation and skewness is calculated according to equation (4),(5),(6):

$$\mu_i = \frac{1}{N}\sum_{j=1}^{n} P_{ij} \quad \ldots (4)$$

$$\sigma_i = \left(\frac{1}{N}\sum_{j=1}^{n}(P_{ij}-\mu_i)^2\right)^{\frac{1}{2}} \quad \ldots (5)$$

$$S_i = \left(\frac{1}{N}\sum_{j=1}^{n}(P_{ij}-\mu_i)^3\right)^{\frac{1}{3}} \quad \ldots (6)$$

Here P_{ij} is the pixel value in thei-th color channel at the j-th image and N is the number of pixels within the image.

B. Texture feature extraction using GLCM

Gray Level Co-Occurrence Matrix (GLCM) is very popular statistical method of extracting textural feature from images. Haralick defines fourteen textural features measured from the co-occurrence matrix to extract the aspects of texture data of images. In this research texture feature are estimated from GLCM matrix. Here we get 1x22 matrix dimensions for texture feature. Some features are contrast, entropy, energy and homogeneity are shown in TABLE 1.

Table 1: Some Features of GLCM

Feature	Formula
Contrast	$\sum_i \sum_j (i-j)^2 p(i,j)$
Entropy	$\sum_i \sum_j p\,(i,j) \log p(i,j)$
Energy	$\sum_i \sum_j p^2(i,j)$
Homogeneity	$\sum_i \sum_j \frac{p(i,j)}{1+\lvert i-j\rvert}$

C. Shape feature extraction using Fourier descriptor

For shape feature extraction of an image we are using Fourier descriptor technique. First we convert an input query image from RGB color space to gray color space. After converting input images we apply canny edge detector to collect edges of the image or boundary of the image. After collecting boundary or edges of an image, calculate Fourier transforms of the boundary through the Fourier descriptors (FD). Fourier descriptors describe shape using Fourier coefficients. Let x(t) and y(t) are the co-ordinates of the boundary and are computed as Z(t) whereas t=0 to maximum

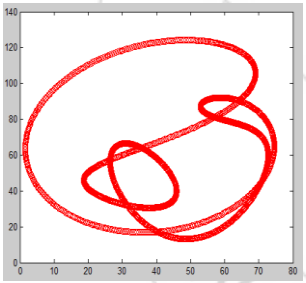
length (L) of boundary co-ordinate. The discrete Fourier transform of Z(t) is calculated using equation (7);

$$U(n) = \frac{1}{N}\sum Z(t)\, exp\left(\frac{-j2n\pi}{N}\right), n = 0,1,\ldots..N-1 \ldots. (7)$$

The coefficients U(n) , n= 0,1,…..N-1, are called Fourier descriptors (FD) of the shape. Hence we get 25 FD's from the shape feature vectors. In matrix form we collect 1x25 features.

(a) Original image

(b) Edge image



(c)Signature of the input image

Figure 2: Shape feature extraction

Figure 2 shows the output of the shape feature extraction. (a) Original image is the query image. (b) Edge image represents the boundary of the query image. (c) Signature of the input query image.

D. Feature Collection

To apply PCA classification method, collect all three features color, texture and shape of an image. In this research we get 9 features of color through color feature extraction, 22 features of texture through texture feature extraction and 25 features of shape through shape feature extraction. After combining them we collect 56 features. All features are combined together for the classification purpose. In matrix form we get 1x9 dimensions for color feature, 1x22 dimensions for texture features and 1x25 dimensions for shape feature. Hence we get total 1x56 features of dimensions. These features will use for classification purpose in the next step.

E. PCA Classification

For feature selection we are using classification method PCA (principal component analysis) algorithm. PCA involves a mathematical procedure that transforms a number of correlated variables into a (smaller) number of uncorrelated variables called principal components. The main objective of the PCA is to reduce the dimensionality of the data set and to identify new meaningful underlying variables.

Let's assume we haveX_i, contain N vectors of size M (= rows of image columns of image) representing a set of images and P represents a pixel values.

$$X_i = [P_1 \ldots\ldots\ldots..P_m]^T, \text{ i = 1…….N} \ldots. (8)$$

Calculate mean of image vector and then set of images are mean centered according to subtract the mean image from every image vector. Let$T_m$represent the mean image.

$$T_m = \frac{1}{M}\sum_{k=1}^{M} X_i \ldots. (9)$$

Let A is the new matrix is constructed by subtracting mean of data from the original data to calculate covariance matrix C.

$$A = X_i - T_m \ldots. (10)$$

Let e's and i's are the eigenvectors and eigenvalues of the covariance matrix C. And this covariance matrix is calculated by multiplying matrix A with its transpose matrix of A.

$$C = AA^T \ldots. (11)$$

The eigenvectors are sorted in descending order with their corresponding eigenvalues. The eigenvector associated with the largest eigenvalue is one that reflects the greatest variance in the image. The number of highest valued eigenvectors is then picked to make an image space from the resultant covariance matrix C.

For testing query image, each image is examined and located its principal features. The CBIR system determines Euclidean distancebetween these principal features and principal features of an image space and chooses that image as a final image whose Euclidean distance is minimum and also classifies the query image to its respective class. Formula to calculate Euclidean distance is:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n}[Z_i - X_i]^2} \ldots. (12)$$

Here Z_i represents test feature data and $X_i$represents original data.

3. Experimental Results

(a) Query Image

(b) Retrieves relevant images for fruit
Fgure 2: Query response for fruit

Figure 3 shows output of the fruit class of the database. (a) Query image represent the test image in the CBIR system and (b) Retrieves relevant images for fruit is the output for query image fruit. The performance rate of this system can be evaluated by the two parameters precision rate and recall rate. The precision rate or value is defined as the ratio of the number of relevant images retrieved to the total number of retrieved images. And recall rate parameter is the ratio of the number of relevant images retrieved to the total number of relevant images. These are the two parameters that improve the performance of our system.

$$\text{Precision} = \frac{\text{No. of relevant images retrieved}}{\text{Total no. of retrieved images}} \quad \ldots (13)$$

$$\text{Recall} = \frac{\text{No. of relevant images retrieved}}{\text{Total no. of relevant images}} \quad \ldots (14)$$

Table 2: Performance Evaluation of our system using precision and recall rate parameter

S. No.	*Classes*	*Precision Rate*	*Recall Rate*
1.	Butterfly	0.67	0.13
2.	Fruit	0.5	0.1
3.	Texture	0.83	0.167
4.	Car	0.83	0.167
5.	Flower	0.67	0.13

The above table 2 represents the precision rate and recall rate for all class of the proposed CBIR system. Precision rate and Recall rate is calculated by the equation (13),(14).

4. Conclusion

In this paper we have proposed an efficient CBIR system where image retrieval method is based on color moment, GLCM features and Fourier descriptor features. To improve retrieval performance we are using PCA classification method so that PCA classifier classifies a new query image to its respective class in the database. PCA based classification system provides a more accurate image classification that infers better and robust data management in various field.

References

[1] Chun.Y, Kim.N, Jang.I,"Content-Based Image Retrieval Using Multiresolution Color and Texture Features," IEEE Transactions On Multimedia, Vol. 10, No. 6, October 2008, pp. 1073-1084.
[2] Durai.R , Duraisamy.V,"A generic approach to content based image retrieval using dct and classification techniques," (IJCSE) International Journal on Computer Science and Engineering ,Vol. 02, No. 06, 2010, pp. 2022-2024.
[3] Arthi.K, Vijayaraghavan.J, "Content Based Image Retrieval Algorithm Using Color Models," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 3, March 2013, pp. 1343-1347.
[4] IBM Research, Almaden. URL: http://www.qbic.almaden.ibm.com
[5] IBM Data Management QBIC URL: http://www.research.ibm.com/topics/popups/deep/manage/html/qbic.html
[6] Trimeche.M, Cheikh.F, Cramariuc.B, and Gabbouj.M, "Content-based Description of Images for Retrieval in Large Databases: MUVIS", X European Signal Processing Conference, Eusipco-2000, vol. 1, September 5-8, 2000.
[7] J. R. Smith, "Integrated spatial and feature image system: Retrieval, analysis and compression " PhD dissertation, Columbia University, New York, 1997
[8] Pentland.A, Picard.R, and Sclaroff.S, "Photobook: Content-Based Manipulation of Image Databases", International Journal of Computer Vision, pp.233-254, 1996.
[9] Bajwa.I, Naweed.M, Asif.M, Hyder.S, "Feature Based Image classification by using Principal component analysis," CIST-Journal of Graphics, Vision and image processing, 2009.
[10] Chen.C, Tseng.Y and Chen.C, "Combination of PCA and Wavelet Transforms for Face Recognition on 2.5D Images," Conf. of Image and Vision Computing 2003, November 2003, pp. 343-347.
[11] Sharma.P, Sharma.D, Goel.N, Kaur.J, "Face Identification Using Wavelet Transform & PCA," Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering, CSEE 2013, pp. 109-113.
[12] Shrivastava.A, Sad.S,"Face Recognition for Different Facial Expressions Using Principal Component analysis," International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 6, July 2014, pp. 326-332.
[13] Pechenizkiy.M ,Tsymbal.A, Puuronen.S, "PCA-based Feature Transformation for Classification: Issues in Medical Diagnostics," IEEE Symposium on Computer-Based Medical Systems, 2004.
[14] Balaji. T, Sumathi. M, "PCA Based Classification of Relational and Identical Features of Remote Sensing Images," International Journal Of Engineering And Computer Science, Volume 3, Issue 7, July 2014, pp. 7221-7228.
[15] http://wang.ist.psu.edu/docs/related/