Research Inventy: International Journal Of Engineering And Science
ISBN: 2319-6483, Issn: 2278-4721, Vol. 1, Issue 8 (November 2012), Pp 18-24
Www.Researchinventy.Com

Feature Extraction for Image Retrieval using Image Mining Techniques

[1]Madhubala Myneni, [2]Dr.M.Seetha
[1,]*Professor In Dept. CSE, AARW, Hydearabad*
[2,]*Professor In Dept. CSE, G.N.I.T. S, Hyderabad*

Abstract: *In this paper feature extraction process is analyzed and a new set of integrated features are proposed for image retrieval. In image retrieval system, the content of an image can be expressed in terms of different features as color, texture and shape. This paper emphasizes on feature extraction algorithms and performance comparison among all algorithms and image mining techniques for converting low level semantic characteristics into high level features. The primitive features are extracted and compared with data set by using various feature extraction algorithms like color histograms, wavelet decomposition and canny algorithms. In this paper feature integration has restricted to five different methodologies of feature Integration: shape only, color only, texture only, color and texture only, and shape, color and texture. It is ascertained that the performance is superior when the image retrieval based on the integrated features, and better results than primitive set.*

Key Words: *Feature Extraction, Feature Integration, Image Retrieval, Image Mining*

I. Introduction

Image Retrieval aims to provide an effective and efficient tool for managing large image databases. Image retrieval and searching is one of the most exciting and fastest growing research areas in the field of digital imaging [2]. The goal of CBIR is to retrieve images from a database that are similar to an image placed as a query. In CBIR, for each image in the database, features are extracted and compared to the features of the query image. A CBIR method typically converts an image into a feature vector representation and matches with the images in the database to find out the most similar images. In various studies different databases have been to compare the study. Content-Based Image Retrieval (CBIR) systems index images using their visual characteristics, such as color, texture and shape, which can be extracted from image itself automatically. The similarity between features was to be calculated using algorithms used by well known CBIR systems such as IBM's QBIC. For each specific feature there is a specific algorithm for extraction and another for matching.

The integration of structure features, which are particularly suitable for the retrieval of manmade objects, and color and texture features, which are geared towards the retrieval of natural images in general. Specifically, the attention was restricted to three different methodologies of feature integration: color and texture, color and shape. Texture and shape results in better performance than using shape, color, and texture individually.

II. Feature Extraction

Feature Extraction is the process of creating a representation, or a transformation from the original data. The images have the primitive features like color, texture, shape, edge, shadows, temporal details etc. The features that were most promising were color, texture and shape/edge. The reasons are color can occur in limited range of set. Hence the picture elements can be compared to these spectra. Texture is defined as a neighbourhood feature as a region or a block. The variation of each pixel with respect to its neighbouring pixels defines texture. Hence the textural details of similar regions can be compared with a texture template. shape/edge is simply a large change in frequency. The three feature descriptors mainly used most frequently during feature extraction are color, texture and shape.

The main method of representing color information of images in Image Retrieval Systems is through color histograms. Quantization in terms of color histograms refers to the process of reducing the number of bins by taking colors that are very similar to each other and putting them in the same bin. There are two types of color histograms, Global color histograms (*GCH*s) and Local color histograms (*LCH*s). A GCH represents one whole image with a single color histogram. An LCH divides an image into fixed blocks and takes the color histogram of each of those blocks. LCHs contain more information about an image but are computationally expensive when comparing images. "The GCH is the traditional method for color based image retrieval. However, it does not include information concerning the color distribution of the regions" of an image. Thus when comparing GCHs

one might not always get a proper result in terms of similarity of images.Texture feature descriptors, extracted through the use of statistical methods, can be classified into two categories according to the order of the statistical function that is utilized: First-Order Texture Features and Second Order Texture Features. First Order Texture Features are extracted exclusively from the information provided by the intensity histograms, thus yield no information about the locations of the pixels. Another term used for First-Order Texture Features is Grey Level Distribution Moments. In contrast, Second-Order Texture Features take the specific position of a pixel relative to another into account. The most popularly used of second-order methods is the Spatial Grey Level Dependency Matrix (SGLDM) method. The method roughly consists of constructing matrices by counting the number of occurrences of pixel pairs of given intensities at a given displacement.

Shape may be defined as the characteristic surface configuration of an object; an outline or contour. Canny edge detection is an optimal smoothing filter given the criteria of detection, localization and minimizing multiple responses to a single edge. This method showed that the optimal filter given these assumptions is a sum of four exponential terms. It also showed that this filter can be well approximated by first-order derivatives of Gaussians. Canny also introduced the notion of non-maximum suppression, which means that given the pre-smoothing filters, edge points are defined as points where the gradient magnitude assumes a local maximum in the gradient direction. Sobel edge detection operations are performed on the data and the processed data is sent back to the computer. The transfer of data is done using parallel port interface operating in bidirectional mode. For estimating image gradients from the input image or a smoothed version of it, different gradient operators can be applied. The simplest approach is to use central differences:

$$L_x(x,y) = -1/2 \cdot L(x-1,y) + 0 \cdot L(x,y) + 1/2 \cdot L(x+1,y).$$
$$L_y(x,y) = -1/2 \cdot L(x,y-1) + 0 \cdot L(x,y) + 1/2 \cdot L(x,y+1).$$

corresponding to the application of the following filter masks to the image data:

$$L_x = \begin{bmatrix} -1/2 & 0 & 1/2 \end{bmatrix} * L \quad \text{and} \quad L_y = \begin{bmatrix} +1/2 \\ 0 \\ -1/2 \end{bmatrix} * L$$

The well-known and earlier Sobel operator is based on the following filters:

$$L_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * L \quad \text{and} \quad L_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * L$$

Given such estimates of first- order derivatives, the gradient magnitude is then computed as:

$$|\nabla L| = \sqrt{L_x^2 + L_y^2}$$

while the gradient orientation can be estimated as

$$\theta = \operatorname{atan2}(L_y, L_x)$$

III. Image Retrieval

In general all Image Retrieval algorithms are based on image primitive features like color, texture and shape. In this paper, the proposal of combinational features is specified to give good performance. On each feature more efficient algorithms are used to retrieve the information from data set.

3.1. Image Retrieval based on Color

In this paper color based image retrieval has performed in two steps. First for extracting color feature information global color histograms has used. To quantize the colors, number of bins are 20. Second step is calculating the distance between bins by using quadratic distance algorithm. The results are the distance from zero the less similar the images are in color similarity.

3.1.1 Color Histograms

Global color histograms extract the color features of images. To quantize the number of bins are 20. This means that colors that are distinct yet similar are assigned to the same bin reducing the number of bins from 256 to 20. This obviously decreases the information content of images, but decreases the time in calculating the color

distance between two histograms. On the other hand keeping the number of bins at 256 gives a more accurate result in terms of color distance. Later on we went back to 256 bins due to some inconsistencies obtained in the color distances between images. There hasn't been any evidence to show which color space generates the best retrieval results, thus the use of this color space did not restrict us an anyway.

3.1.2 Quadratic Distance Algorithm

The equation we used in deriving the distance between two color histograms is the quadratic distance metric:

$$d^2(Q,I) = \left(H_Q - H_I\right)^t A\left(H_Q - H_I\right)$$

The equation consists of three terms. The derivation of each of these terms will be explained in the following sections. The first term consists of the difference between two color histograms; or more precisely the difference in the number of pixels in each bin. This term is obviously a vector since it consists of one row. The number of columns in this vector is the number of bins in a histogram. The third term is the transpose of that vector. The middle term is the similarity matrix. The final result *d* represents the color distance between two images. The closer the distance is to zero the closer the images are in color similarity. The further the distance from zero the less similar the images are in color similarity.

3.1.3 Similarity Matrix

As can be seen from the color histograms of two images *Q* and *I*, the color patterns observed in the color bar are totally different. A simple distance metric involving the subtraction of the number of pixels in the 1st bin of one histogram from the 1st bin of another histogram and so on is not adequate. This metric is referred to as a minkowski-form distance metric, which only compares the "same bins between color histograms ".
This is the main reason for using the quadratic distance metric. More precisely it is the middle term of the equation or similarity matrix ***A*** that helps us overcome the problem of different color maps. The similarity matrix is obtained through a complex algorithm:

$$a_{q,i} = 1 - \frac{\left[\left(v_q - v_i\right)^2 + \left(s_q \cos\left(h_q\right) - s_i \cos\left(h_i\right)\right)^2 + \left(s_q \sin\left(h_q\right) - s_i \sin\left(h_i\right)\right)^2\right]^{\frac{1}{2}}}{\sqrt{5}}$$

which basically compares one color bin of $\mathbf{H_Q}$ with all those of $\mathbf{H_I}$ to try and find out which color bin is the most similar.

3.2 Image Retrieval based on Texture

The texture based image retrieval was performed in three steps. First for extracting statistical texture information Pyramid-structured wavelet transform has used. This decomposition has been done in five levels. Second step is calculating energy levels on each decomposition level. Third step is calculating euclidian distance between query image and database images. The top most five images from the list are displayed as query result.

3.2.1. Pyramid-Structured Wavelet Transform

This transformation technique is suitable for signals consisting of components with information concentrated in lower frequency channels. Due to the innate image properties that allows for most information to exist in lower sub-bands, the pyramid-structured wavelet transform is highly sufficient. Using the pyramid-structured wavelet transform, the texture image is decomposed into four sub images, in low-low, low-high, high-low and high-high sub-bands. At this point, the energy level of each sub-band is calculated .This is first level decomposition. Using the low-low sub-band for further decomposition, we reached fifth level decomposition, for our project. The reason for this is the basic assumption that the energy of an image is concentrated in the low-low band. For this reason the wavelet function used is the daubechies wavelet.
For this reason, it is mostly suitable for signals consisting of components with information concentrated in lower frequency channels. Due to the innate image properties that allows for most information to exist in lower sub-bands, the pyramid-structured wavelet transform is highly sufficient.

3.2.2 Energy Level

Energy Level Algorithm:

- Decompose the image into *four* sub-images
- Calculate the energy of all decomposed images at the same scale, using [2]:

$$E = \frac{1}{MN} \sum_{i=1}^{m} \sum_{j=1}^{n} \left| X(i, j) \right|$$

where ***M*** and ***N*** are the dimensions of the image, and ***X*** is the intensity of the pixel located at row ***i*** and column ***j*** in the image map.

- Repeat from step 1 for the low-low sub-band image, until index ***ind*** is equal to 5. Increment ***ind***.

Using the above algorithm, the energy levels of the sub-bands were calculated, and further decomposition of the low-low sub-band image. This is repeated five times, to reach fifth level decomposition. These energy level values are stored to be used in the Euclidean distance algorithm.

3.2.3 Euclidean Distance

euclidean distance algorithm:

- Decompose query image.
- Get the energies of the first dominant ***k*** channels.
- For image ***i*** in the database obtain the ***k*** energies.
- Calculate the euclidean distance between the two sets of energies, using [2]:

$$D_i = \sum_{k=1}^{k} \left(x_k - y_{i,k} \right)^2$$

- Increment ***i***. Repeat from step 3.

Using the above algorithm, the query image is searched for in the image database. The euclidean distance is calculated between the query image and every image in the database. This process is repeated until all the images in the database have been compared with the query image. Upon completion of the euclidean distance algorithm, we have an array of euclidean distances, which is then sorted. The five topmost images are then displayed as a result of the texture search.

3.3 Image Retrieval based on Shape

In this paper shape based image retrieval has performed in two steps. First for extracting edge feature canny edge detection algorithm and sobel edge detection algorithm are used. Second step is calculating euclidian distance between query image and database images. The top most five images from the list are displayed as query result.

3.3.1. Canny Edge Detection Algorithm

The canny edge detection algorithm was used to detect a wide range of edges in images. The stages of the canny algorithm include the noise reduction, finding the intensity gradient of the image, non-maximum suppression, tracing edges through the image and hysteresis threshold, differential geometric formulation of the canny edge detector

IV. Feature Integration

All the extracted features are integrated to get the final extracted image as result. Every block has a similarity measure for each of the features. Hence after the feature extraction process each instance (block) is a sequence of 1s (YES) and 0s (NO) of length equal to the number of features extracted. Combining these extracted features is synonymous to forming rules. One rule that combines the three features is color & edge | textures, which means color AND edge OR texture. Depending on the features used and the domain, the rules vary. If a particular feature is very accurate for a domain, then the rule will assign the class label as YES (1) (1 in the table on the left). For those instances when I Class is not certain the class label is 2. This denotes uncertain regions that may or may not be existed. The same rule used during the training phase is also used in the testing phase.

Table1

Rules for Feature Integration

Rule Name	Color	Texture	shape
Class-1	1	0	1
Class-2	0	0	0
Class-3	1	1	0
Class-4	1	0	0

If there are 3 features, for example, the above table shows a part of a set of rules that could be used. The first and

third rules say that color along with texture or edge conclusively determines that query image is present in that block. The second rule says that when none of the features is 1 then query image is absent for sure. Fourth rule states that color on its own is uncertain in determining the presence of query image.

V. Results and Discussions

The image database has used to retrieve the relevant images based on query image. The test image database contains 200 images of 10 categories like structured and unstructured, sports images, missile images etc. Image retrieval was performed on combinational feature set of primitive features like color, texture and shape. Results are obtained from class 1 feature set of primitive feature color, class 2 feature set of integrated feature color and shape and class 3 feature set of integrated feature texture and shape. For single query missile image the results shown in the following figures.

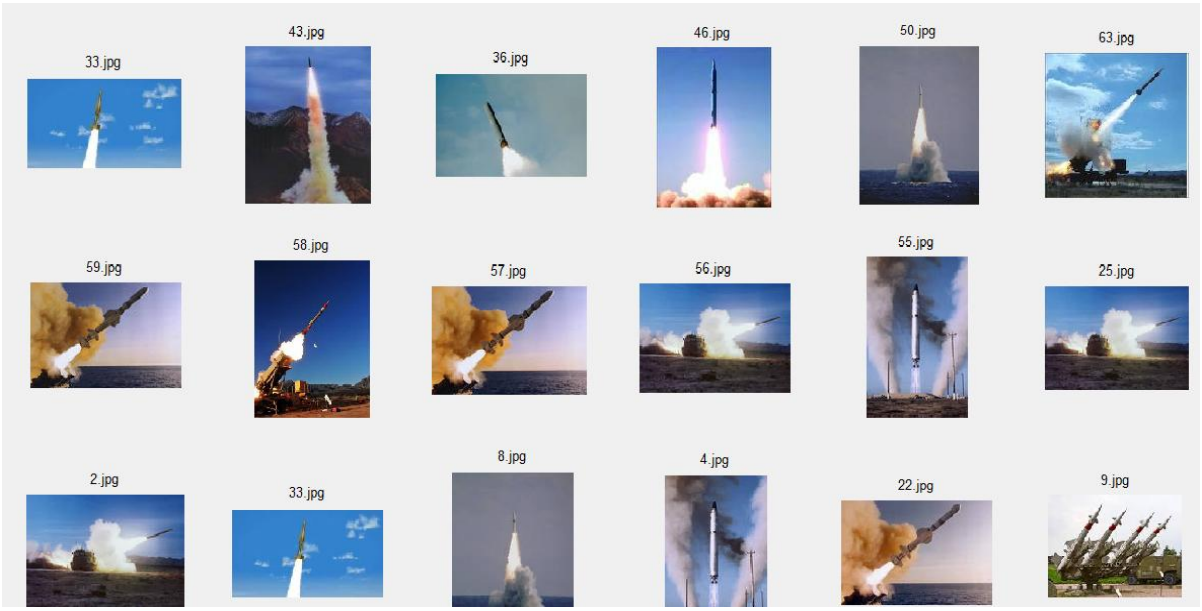
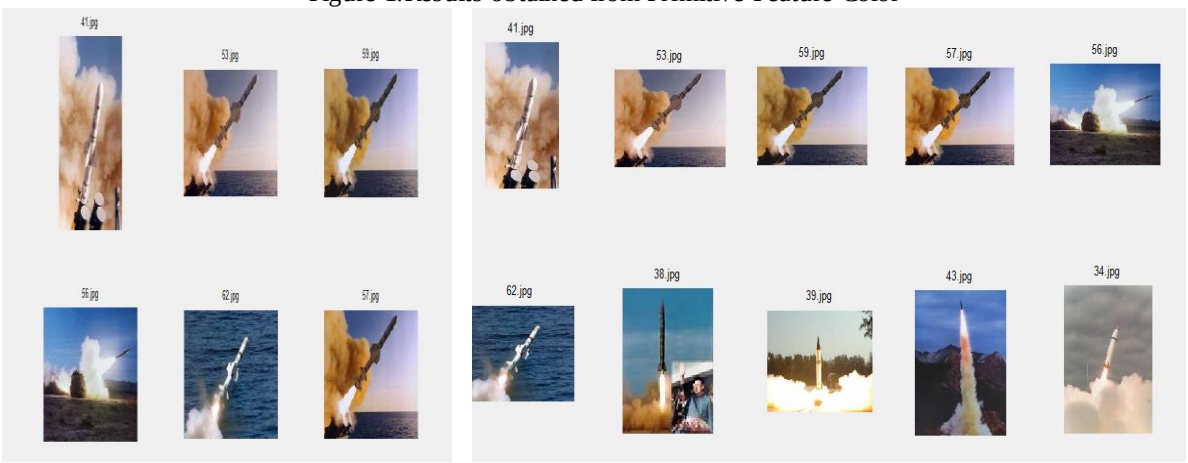


Figure 1.Results obtained from Primitive Feature Color



(a) (b)

Figure 2. Results obtained from Integrated Feature Extraction (a)color and texture ,(b)color and shape

Based on commonly used performance measures in information retrieval, two statistical measures were computed to assess system performance namely Recall and Precision. For good retrieval system ideal values for recall is 1 and precision is of low value. The table 2 will give the performance evaluation of image retrieval based on primitive feature color on 10 image databases.

Table 2
Performance Analysis based on Primitive Feature (Color)

Query database	Retrieved Images	Relevant Images	Relevent Images retrieved	Precision	Recall
Img1	9	5	5	0.444	1
Img2	9	7	6	0.666	0.857
Img3	9	5	5	0.555	1
Img4	9	6	6	0.667	1
Img5	9	6	6	0.667	1
Img6	9	5	4	0.444	0.8
Img7	9	8	6	0.667	0.75

The table 2 will give the performance evaluation of integrated features of different classes like color and texture, color and shape.

Table 3
Performance analysis based on integrated features of color and texture and color and shape

Query database	**Integrated features of color and texture**		**Integrated features of color and shape**	
	Precision	Recall	Precision	Recall
Img1	0.833	1	1	0.8
Img2	0.833	0.714	0.75	0.428
Img3	0.833	0.8	0.75	0.6
Img4	1	1	1	0.667
Img5	0.833	0.833	1	0.667
Img6	0.667	0.8	1	0.8
Img7	0.83	0.625	0.75	0.375

The performance analysis of image retrieval system is based on primitive feature color and integrated features color and texture and color and shape. By taking 10 query images from different databases, the results are analysed based on number of relevant images retrieved and total relevant images existed in the data base. The performance of image retrieval system of integrated feature is more when compared with primitive feature extraction.

VI. Conclusions and Future Enhancements

This paper elucidates the potentials of extraction of features of the image using color, texture and shape for retrieving the images from the specific image databases. The images are retrieved from the given database of images by giving the query image**.** These results are based on various digital images of dataset. The performance of the image retrieval was assessed using the parameters recall rate and precision. It was ascertained that the recall rate and precision are high when the image retrieval was based on the feature integration on all the three features the color, texture and shape than primitive features alone. The work can be extended further on huge data bases for retrieving relevant images to obtain objects using different combination of weight for color and texture and shape features.

References

[1]. A.W.M. Smeulders, et. al. Content-based imageretrieval at the end of the early years, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12,2000, pp. 1349.

[2]. Castelli, V. and Bergman, L. D., Image Databases: Search and Retrieval of Digital Imagery, 2001, John Wiley & Sons, Inc.

[3]. Craig Nevill-Manning (Google) and Tim Mayer (Fast). *Anatomy of a Search Engine, Search Engine Strategies Conference*, 2002.

[4]. Daubechies, I., Ten Lectures on Wavelets, *Capital City Press, Montpelier, Vermon.t*,1992.

[5]. Drimbarean A. and Whelan P.F. Experiments in color texture analysis, P *attern Recognition Letters*, 22:2001, 1161-1167.

[6]. Hedman, K.; Stilla, U.; Lisini, G.; Gamba, P. , Road Network Extraction in VHR SAR Images of Urban and Suburban Areas by Means of Class-Aided Feature-Level Fusion, *Geoscience and Remote Sensing, IEEE Transactions* , vol.48, no.3, pp.1294-1296, March 2010.

[7]. Hui Kong; Audibert, J.-Y.; Ponce, J. , General Road Detection From a Single Image, *Image Processing, IEEE Transactions* , vol.19, no.8, pp.2211-2220, Aug. 2010.

[8]. Jalal, A. A Fuzzy Model for Road Identification in Satellite Images. *Proceedings of the 2006 International Conference on Image Processing, Computer Vision, & Pattern Recognition, Las Vegas, Nevada, USA*, 2006.

[9]. Nock, R.; Nielsen, F.,Statistical region merging, pattern analysis and machine intelligence, *IEEE Transactions on* , vol.26, no.11, pp.1452-1458, Nov. 2004.

[10]. Shou-Bin Dong and Yi-Ming Yang. Hierarchical Web Image Classification by Multi-level Features, *Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing*, 2002

[11]. Smith, J., Color for Image Retrieval, *Image Databases: Search and Retrieval of Digital Imagery, John Wiley & Sons, New York*. 2001.

[12]. Tomoko Tateyama, Zensho Nakao, Xian Yan Zeng, Yen-Wei Chen, Segmentation of High Resolution Satellite Images by Direction and Morphological Filters, *his pp.482-487, Fourth International Conference on Hybrid Intelligent Systems* (HIS'04), 2004.

[13]. Tuncer, O. ,Fully Automatic Road Network Extraction from Satellite Images, *Recent Advances in Space Technologies, 2007. RAST '07. 3rd International Conference on* , vol., no., pp.708-714, 14-16 June 2007.

Biography

M.Madhu Bala.
She is doing her Ph.D in Computer Science and Engineering in the area of image mining at Jawaharlal Nehru Technological University, Hyderabad. Her research interests are DataMining, Image Analysis, Information Retrieval Systems.. She holds the Life Membership of ISTE and CSI.

Dr. M.Seetha.

She had completed Ph.D in Computer Science and Engineering in the area of image processing in December 2007 from Jawaharlal Nehru Technological University, Hyderabad and M. S. from B I T S, Pilani in 1999. Her research interest includes image processing, neural networks, computer networks, artificial intelligence and data mining. She had about 10 papers published in refereed journals and more than 50 papers in the proceedings of National/International Conference and Symposiums. She was the recipient of the **AICTE Career Award for Young Teachers (CAYT)** in FEB, 2009, and received the grant upto **10.5 lakhs** over a period of three years by AICTE, INDIA. She was a reviewer for various International Journals/Conferences. She holds the Life Membership of ISTE, IETE and CSI.