Extracting The Features of Web Images: Effective Methods of Image Mining Techniques

R . Lakshman Naik, Dr . B . Manjula
Department of computer science
Kakatiya University, Warangal, A.P., India

Abstract: **Due to the digitization of data and advances in technology, it has become extremely easy to obtain and store large quantities of data, particularly Multimedia data. Mining Image data is the one of the essential features in this present scenario since image data plays vital role in every aspect of the system such as business for marketing, hospital for surgery, engineering for construction, Web for publication and so on. Feature selection and extraction is the pre-processing step of Image Mining. Obviously this is a critical step in the entire scenario of Image Mining. To extract patterns and derive knowledge from large collections of images, deals mainly with identification and extraction of unique features for a particular domain. There are various features available in the image, to identify the best features and thereby extract relevant information from the images. In this paper, we have proposed three effective methods of image mining techniques (QBIR, BTA, and clustering algorithm) for extracting the features of the web image.**

Keywords: **Image mining, clustering, Extract, Query, content-based, Database**

1. INTRODUCTION

The rapid increase in the digital images for multimedia system is used for hiding this data collection that is potentially used in the wide range of applications like business, medical, Geographical Information System (GIS), remote sensing, and Crime Prevention, Military, Home Prevention and World Wide Web (WWW) and so on. If we analysis these data, which can reveal useful information to the human users.

Image processing is any form of signal processing where the input can be a photograph or a video frame and the output may be either an image or a set of parameters related to the image. An image retrieval system is a system which allows us to browse, search and retrieve the images. Query by Image content (QBIC) Retrieval [6, 7] is the process of retrieving the desired query image from a huge number of databases based on the contents of the image.

In recent years, researcher has been showing interest in developing effective methods for content based image clustering and retrieval. This interest has been motivated by the need to efficiently manage large image databases and efficiently run image retrievals to get the best results without exhaustively searching the global database each time. This leads to huge savings in time and money, especially in fields where the bulk of working databases are image files or any kind of media whose contents cannot be adequately described by simple keywords or short texts. Image mining deals with the extraction of knowledge, image data relationship, or other patterns not explicitly stored in the images. It uses methods from computer vision, image processing, image retrieval, data mining, machine learning, database, and artificial intelligence. Rule mining has been applied to large image databases. There are two main approaches. The first approach is to mine from large collections of images alone and the second approach is to mine from the combined collections of images and associated alphanumeric data [1]. Clustering is a method of grouping data objects into different groups, such that similar data objects belong to the same group and dissimilar data objects to different clusters. Image clustering consists of two steps the former is feature extraction and second part is grouping [2].

In this paper we will study about a data mining approach to cluster the images based on colours. And further the colour features obtained are used to cluster the images. Colour features extraction is done using the block truncation algorithm and clustering using the k-means algorithm. Query-By Image Content retrieval

systems aims at searching image databases for specific images that are similar to a given query image.

2. QUERY BY IMAGE CONTENT

Query by image content (QBIC) Retrieval System is also called as Content-Based Image Retrieval (CBIR) systems and Content-Based Visual Information Retrieval (CBVIR) system, utilize low level query image feature as identifying similarity between a query image and the image database. "content-based" means that the search will analyze the actual contents of the image rather than the metadata such as keywords, tags, and/or descriptions associated with the image. Image contents are plays significant role for image retrieval. Each image has three contents such as: colours, texture and shape features. Colour and texture both plays important image visual features used in Query-By Image Content Retrieval to improve results.

A typical QBIC system automatically extract visual attributes like colour, shape, texture and spatial information of each image in the database based on its pixel values and stores them in to a dissimilar database within the system called feature database. The feature data for each of the visual attributes of each image is very much smaller in size compared to the image data. The feature database contains an abstraction of the images in the image database; each image is represented by compact illustration of its contents like color, texture, shape and spatial information in the form of a fixed length real-valued multi-component feature vectors or signature.

Query by example is a query technique that involves providing the QBIC system with an example image that it will then base its search upon. Options for providing example images to the system include: pre-existing image may be supplied by the user or chosen from a random set and the user draws a rough approximation of the image they are looking for. For example with blobs of colour or general shapes. This query technique removes the difficulties that can arise when trying to describe images with words. The figure.1 shows the General model of QBIC.

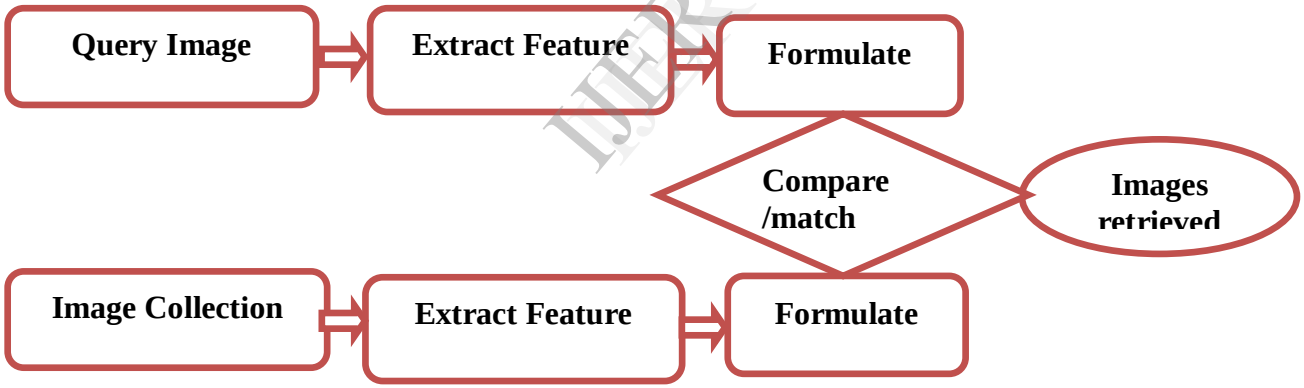


Figure 1: General model of QBIC

A. Data gathering

By using Internet spider program that can collect webs automatically to interview Internet and do the gathering of the Images on the web site, then it will go over all the other webs through the URL, repeating this process and collecting all the images it has reviewed into the server.

B. Extract feature database

Using index system program do analysis for the collected images and extract the feature information. At this time, the features that use widely involve low-level features such as color; texture and so on, the middle-level features such as shape.

C. Searching in the Database

System extract the feature of image that waits for search when user input the image sample that need search, then the search engine will search the suitable feature from the database and calculate the similar distance, then find some related webs and images with the lowest similar distance.

D. Process and index the results

Index the image obtained from searching due to the similarity of features, and then returns the retrieval images to the user and allows the user select. If the user is not pleased with the searching result, he can re-retrieval the image again, and searches database again.

3. FEATURE EXTRACTION

The properties of colour, texture, and shape have broad, intuitive applicability. Feature (content) extraction is the basis of content-based image retrieval. Corresponding features are computed for all objects and full scenes and stored for use in subsequent queries. The computed features are: colour, shape and texture.

A. Colour features

The color feature extraction procedure includes color image segmentation. Examining images based on the colors they contain is one of the most widely used techniques because it does not depend on image size or orientation. There are lots of techniques available for retrieving images on the basis of color similarity from image database. Each image included to the collection is analyzed to compute a color histogram which shows the proportion of pixels of each color within the image. The color histogram for each image is then stored in the database [3].

B. Texture features

Texture features are based on modified versions of the coarseness, contrast, and directionality features [4]. The coarseness feature helps measure the scale of the texture. The contrast feature describes the vividness of the pattern, and is a function of the variance of the gray-level histogram. The directionality feature describes whether or not the image has a favoured direction or whether it is isotropic. Textures characterization is useful in distinguishing between areas of images with similar colours (such as sky, sea, or water, grass).

C. Shape features

One of the most challenging aspects to content based image retrieval is retrieval by shape. The image is converting into binary. Polygonal approximation that uses straight-line, Bezier curve and BSpline are applied. As a result the image is presented as a set of straight lines, arcs and curves [1]. Shape representation is significant concern both in object recognition and classification. That means an image has to be segmented before extracting most shape features.

D. Other types of primitive features

Several other types of image feature have been proposed as a basis for CBIR. Most of these rely on complex transformations of pixel intensities which have no obvious counterpart in any human description of an image. Most such techniques aim to extract features which reflect some aspect of image similarity which a human subject can perceive, even if he or she finds it difficult to describe. The well-researched technique of this kind uses the wavelet transform to model an image at several different resolutions. Promising retrieval results have been reported by matching wavelet features computed from query and stored images [9, 10]. Another method giving interesting results is retrieval by appearance. Two versions of this method have been developed, one for whole-image matching and one for matching selected parts of an image. The part-image technique involves filtering the image with Gaussian derivatives at multiple scales [11], and then computing differential invariants; the whole-image technique uses distributions of local curvature and phase [12].

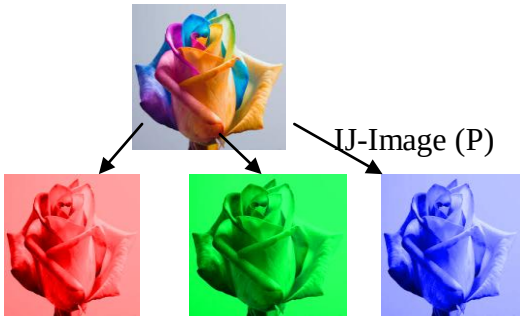
The advantage of all these techniques is that they can describe an image at varying levels of detail (useful in natural scenes where the objects of

interest may appear in a variety of guises), and avoid the need to segment the image into regions of interest before shape descriptors can be computed. Despite recent advances in techniques for image segmentation [13], this remains a troublesome problem.

4. BLOCK TRUNCATION ALGORITHM

Block Truncation algorithm is a type of lossy image compression technique for greyscale images. It divides the original images into blocks and then uses a quantise to reduce the number of grey levels in each block whilst maintaining the same mean and standard deviation. Steps in Block Truncation Coding Algorithm:

Step1: Split the image into Red, Green, Blue Components



Red (R(P)) Green (G(P)) Blue (B(P))

Step2: Find the average of each component

$$red\ Avg = \frac{sum of\ all\ red\ Pixel\ in the image\ R(P)}{no.of\ Pixels\ in image\ P}$$

$$green\ Avg = \frac{sum of\ all\ red\ Pixel\ in the image\ G(P)}{no.of\ Pixels\ in image\ P}$$

$$blue\ Avg = \frac{sum of\ all\ red\ Pixel\ in the image\ B(P)}{no.of\ Pixels\ in image\ P}$$

Where R (P) = Red component pixels,

G (P) = Green component pixels,

B (P) = Blue component pixels,

P =No. of pixels in the image.

After calculating the mean values of Red, Blue and Green components, the values are to be compared with each other in order to find the maximum value of the components. For e.g. if the value of Red component is High than the rest of the two, then we can conclude that the respective image is Red Intensity oriented image and which can be clustered into Red Group of Images. Whenever the query image is given, calculate the RGB components average values. Then compare this with the stored values.

Step3: Split every component image to obtain RH, RL, GH, GL, BH and BL images

RH is obtained by taking only red component of all pixels in the image which are above red average and RL is obtained by taking only red component of all pixels in the image which are below red average. Similarly GH, GL, BH and BL can be obtained.

Step4: Apply colour moments to each splited component, i.e. RH, RL, GH, GL, BH and BL images.

And for the Color moments, that can be used differentiate images based on their features of color, probability distributions is used. 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 in which P_{ij}^{k} is the value of the k[th] color component of the ij-image pixel and P is the height of the image, and Q is the width of the image. They are Mean (E_k), Standard deviation (SD_k) and Skewness (S_k).

Mean can be understood as the average color value in the image.

$$E_k = \frac{1}{PQ}\sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}^{k}$$

The standard deviation is the square root of the variance of the distribution.

$$SD_k = \sqrt{\left(\frac{1}{PQ}\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{ij}^{k} - E_k)^2\right)}$$

Skewness can be understood as a measure of the degree of asymmetry in the distribution.

$$S_k = \left(\frac{1}{PQ}\sum_{i=1}^{n}\sum_{j=1}^{n}(P_{ij}^{k} - E_k)^3\right)^{\frac{1}{3}}$$

Step5: Apply clustering algorithm to find the clusters.

5. CLUSTERING ALGORITHM

K-means algorithm is one of the most widely used clustering algorithms [8] in spatial clustering analysis. It is easy and efficient. K-means is one of the simplest unsupervised learning algorithms in which each point is assigned to only one particular cluster. The procedure of K-mean algorithm consists of the following steps:

Step 1: Set the number of cluster k

Step 2: Determine the centroid coordinate

Step 3: Determine the distance of each object to the centroids

Step 4: Group the object based on minimum distance

Step 5: Continue from step 2, until convergence that is no object move from one group to another.

6. CONCLUSION

Large online image collections are becoming more common and methods to manage, organize, and retrieve images from this collected database need an effective method. These low-level features are extracted directly from digital representations of the image and do not necessarily match the human perception of visual semantics. We proposed a framework of unsupervised clustering of images based on the color feature of image.

The main aim of this paper is to get the knowledge of about the image retrieval system. Here Block Truncation Algorithm is used to extract features for image dataset and K-Means clustering algorithm is conducted to group the image dataset into various clusters. The Query-By Image content retrieval systems are used to matches the image from Query Image Features with image Feature Database. Our future work is to apply these algorithms to various fields for retrieving the images and to evaluate the performance of these algorithms

REFERENCES

1. Peter Stanchev, "USING IMAGE MINING FOR IMAGE RETRIEVAL", IASTED,
"Computer Science and Technology", May 19-21, 2003 Cancum, Mexico, 214-218.
2. Dr. Sanjay Silakari, Dr. Mahesh Motwani and Manish Maheshwari, "Color Image Clustering using Block Truncation Algorithm" IJCSI,Vol. 4, No. 2, 2009.
3. Jayant Mishra, Anubhav Sharma and Kapil Chaturvedi "An Unsupervised Cluster-based Image Retrieval Algorithm using Relevance Feedback", IJMIT, Vol.3, No.2, May 2011.
4. 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", IBM Research Division, Almaden Research Center K54/802, 650 Harry Road, San Jose, CA 95120-6099
5. A.Kannan, Dr. V. Mohan, Dr. N. Anbazhagan, "Image Clustering and Retrieval using Image Mining Techniques", 2010 IEEE International Conference on Computational Intelligence and Computing Research.
6. Wayne Niblack, et al(1993) "The QBIC Project: Querying Images by Content, Using Color, Texture, and Shape", Storage and Retrieval for Image and Video Databases (SPIE) 1993: 173-187.
7. Ji Zhang Wynne Hsu Mong Li Lee, "Image Mining: Trends and Developments".
8. Han, M.Kamber, "Data Mining concepts and Techniques", Morgan Kaufmann Publishers, 2002
9. Jacobs, C. E. Et al (1995) "Fast Multiresolution Image Querying" Proceedings of SIGGRAPH 95, Los Angeles, CA (ACM SIGGRAPH Annual Conference Series, 1995), 277-286.
10. Liang, K C and Kuo, C C J (1998) "Implementation and performance evaluation of a progressive image retrieval system" in Storage and Retrieval for Image
and Video Databases VI (Sethi, I K and Jain, R C, eds), Proc SPIE 3312, 37-48
11. Ravela, S and Manmatha, R (1998a) "Retrieving images by appearance" in Proceedings of IEEE International Conference on Computer Vision (IICV98), Bombay, India, 608-613
12. Ravela, S and Manmatha, R (1998b) "On computing global similarity in images" in Proceedings of IEEE Workshop on Applications of Computer Vision (WACV98), Princeton, NJ, 82-87
13. Campbell, N W et al (1997) "Interpreting Image Databases by Region Classification" Pattern Recognition 30(4), 555-563
14. M.Stricker and M.Orengo, "Similarity of color images",Storage and Retrieval for Image and Video Databases III (SPIE) 1995: 381-392