Improve Web Mining Search Engine using 3d Histogram

Savita, Brijesh vishwakarma
Computer Science and Engineering,
Bansal Institute of Engineering and Technology, Lucknow, India

Abstract: The retrieval principle of CBIR systems is based on visual features such as colour, texture, and shape or the semantic meaning of the images. To enhance the retrieval speed, most CBIR systems pre-process the images stored in the database. This is because feature extraction algorithms are often computationally expensive. If images are to be retrieved from the World-Wide-Web (WWW), the raw images have to be downloaded and processed in real time. In this case, the feature extraction speed becomes crucial. Ideally, systems should only use those feature extraction algorithms that are most suited for analyzing the visual features that capture the common relationship between the images in hand. In this thesis, a statistical discriminate analysis based feature selection framework is proposed. Such a framework is able to select the most appropriate visual feature extraction algorithms by using relevance feedback only on the user labeled samples. The idea is that a smaller image sample group is used to analyze the appropriateness of each visual feature, and only the selected features will be used for image comparison and ranking. As the number of features is less, an improvement in the speed of retrieval is achieved.

Keywords: CBIR, Histogram, RGB, Digital Image

1. Introduction:

Recent advances of the technology in digital imaging, broadband networking and digital storage devices make it possible to easily generate, transmit, manipulate and store large numbers of digital images and documents. As a result, image databases have become widespread in many areas such as art gallery and museum management, architectural and engineering design, interior design, remote sensing and management of earth resources, geographic information systems, medical imaging, scientific database management systems, weather forecasting, fabric and fashion design, trade-mark and copyright database management, law enforcement, criminal investigation, picture archiving and communication systems. Furthermore, the rapid growth of the World Wide Web has led to the formation of a very large but disorganized, publicly available image collection. Recent studies show that there are 180 million digital images on publicly indexable Web and millions of new images are being produced every day. Thus, efficient image retrieval from digital image collections have been of great interest over the last decade and several systems have been developed for research and commercial purposes.

Content-based Image Retrieval (CBIR), in itself, has proved to be a high potential field of research with ever-increasing demands of higher performance, effective retrieval and the need of incorporating greater machine intelligence into the process. Therefore, research activity in the subject of CBIR has increased significantly over the past decade. A considerable amount of data, especially in fields like medical imagery, remote sensing, multimedia etc. is available to our disposal. With the advancements in technology, more and more data is generated which if used properly can prove to be a great source of information related to the respective domains. It would prove to be highly useful to use CBIR systems in such fields to extract the information available in these huge data repositories. Basically, the web consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks. Recent research on mining, multi-types of data is termed as multimedia data mining. This line of research is yet to receive proper attention and most of the efforts on web content mining. There are a number of text search engines on the web and incidentally, the site hosting them, are amongst the business sites. However, searching for multimedia content is not an easy task because the multimedia data as opposed to text needs many stages of pre processing to yield indices relevant for querying. Since an image or a video sequence can be interpreted in numerous ways, there is no commonly agreed upon vocabulary. Thus, the strategy of manually assigning a set of labels to a multimedia data, storing it and matching the stored label with a query will not be effective. Besides, the large volume of video data makes any assignment of text labels a massively labor intensive effort.

Recently information retrieval for multimedia content has become an important research area. Content-based retrieval in multimedia is a challenging task since multimedia data needs detailed interpretation from pixel values. Automated analysis calculates statistics, which can be approximately correlated to the content features. This is useful as it provides information without human interaction. There is a great need to extract semantic indices for making the content-based retrieval system serviceable to the user. Though extracting all such indices might not be possible, there is a great scope for furnishing the semantic indices with a certain well-established structure.

2. Related Work:

Content-based image retrieval (CBIR), as we see it today, is any technology that in principle helps to organize digital picture archives by their visual content. By this definition, anything ranging from an image similarity function to a robust image annotation engine falls under the purview of CBIR. This characterization of CBIR as a field of study places it at a unique juncture within the scientific community. While we witness continued effort in solving the fundamental open problem of robust image understanding, we also see people from different fields, such as, computer vision, machine learning, information retrieval, human-computer interaction, database systems, Web and data mining, information theory, statistics, and psychology contributing and becoming part of the CBIR community [1]. Moreover, a lateral bridging of gaps between some of these research communities is being gradually brought about as a by-product of such contributions,

the impact of which can potentially go beyond CBIR. Again, what we see today as a few cross-field publications may very well spring into new fields of study in the foreseeable future.

Amidst such marriages of fields, it is important to recognize the shortcomings of CBIR as a real-world technology. One problem with all current approaches is the reliance on visual similarity for judging semantic similarity, which may be problematic due to the semantic gap [2] between low-level content and higher-level concepts. While this intrinsic difficulty in solving the core problem cannot be denied, we believe that the current state-of-the-art in CBIR holds enough promise and maturity to be useful for real-world applications if aggressive attempts are made.

The video-sharing and distribution forum YouTube has also brought in a new revolution in multimedia usage. Of late, there is renewed interest in the media about potential real-world applications of CBIR and image analysis technologies, as evidenced by publications in Scintific American [4], Discovery News [5] and on [6].

We envision that image retrieval will enjoy a success story in the coming years. We also sense a paradigm shift in the goals of the next-generation CBIR researchers. The need of the hour is to establish how this technology can reach out to the common man in the way text retrieval techniques have. Methods for visual similarity, or even semantic similarity (if ever perfected), will remain techniques for building systems. What the average end-user can hope to gain from using such a system is a different question altogether.
Comprehensive surveys exist on the topic of CBIR [7, 8, 9], all of which deal primarily with work prior to the year 2000. Surveys also exist on closely related topics such as relevance feedback [10], high-dimensional indexing of multimedia data [11], face recognition [10] (useful for face-based image retrieval), applications of CBIR to medicine, and applications to art and cultural imaging [12]. In our current survey, we restrict the discussion to image-related research only.
One of the reasons for writing this survey is that CBIR, as a field, has grown tremendously after the year 2000 in terms of the people involved and the papers published. Lateral growth has also occurred in terms of the associated research questions addressed, spanning various fields. To validate the hypothesis about growth in publications, we conducted a simple exercise. We searched for publications containing the phrases “Image Retrieval” using Google Scholar [l3] and the digital libraries of ACM, IEEE, and Springer, within each year from 1995 to 2005. In order to account for: (a) the growth of research in computer science as a whole, and (b) Google’s yearly variations in indexing publications, the Google Scholar results were normalized using the publication count for the word “computer” for that year. A plot on another young and fast-growing field within pattern recognition, support vector machines (SVMs), was generated in a similar manner for comparison. Not surprisingly, the graph indicates similar growth patterns for both fields, although SVM has had faster growth. These trends indicate, given the implicit assumptions, a roughly exponential growth in interest in image retrieval and closely related topics. We also observe particularly strong growth over the last five years, spanning new techniques, support systems, and application domains.

3. Methodology:
The two major issues of image retrieval systems are indexing and retrieval. Indexing is to extract the features of the image without losing any useful information. The extracted features are then organized in a specific form and stored in an index file. Retrieval means compare the indexed data with the query data and get the most relevant image. The following three techniques are used in our system:

3.1Color Analysis:
This method analyses the color composition of the image. A RGB model is used to represent all colors. It is a 3-dimenisonal model and the color is represented by the magnitude of the three vectors: Red (R), Green (G) and Blue (B). The magnitude of each vector is from 0 to 255. Thus totally 2563 = 16.7 million types of colors can be represented. To reduce the size of the index file, each color vector is subdivided into sections and different color bins are formed. If the number of sectors is 4 then the number of color bins will be 43 = 64. If the images in the database are quite different, then using less number of color bins is already enough. The RGB value of each pixel is read and mapped to the corresponding color bin. After scanning the whole image, a distribution of color histogram is generated. The normalized color histogram is stored in the index file.

3.2 Image Retrieval Method:
To retrieve an image form the database, we first analyze the sample image inputted by the user using the above analysis and form the sample index. Then we read data from the index file and calculate the similarity value between the stored image and the input image based on absolute difference or generalized similarity matrix. The image with the highest similarity is then selected.

3.2.1 Absolute Difference
This is the most straightforward method. To compare two images, we compute the similarity value S_D as follows

$$S_D(X,Y) = \sum_{k=1}^{N} |X_k - Y_k|$$

where Xk and Yk are the percentage of pixels of the corresponding color/edge bin k in image X: and image Y respectively. N denotes the number of colour/edge bins. Obviously the larger the value of &(X, Y), the less similar the two images.

3.2.2 Generalized Similarity Matrix:
The absolute difference method does not cater the relationship among different color bins. If two colors which look similar perceptually but fall into different color bins, they will be considered as totally different in the calculation of the similarity value. Consequently the retrieval result will be worse than expected. To overcome this weakness the similarity matrix A = [a(i, j)] is introduced. The values assigned in A specify the weighting relationship among different color bins and are calculated as follows:

$$a(i, j) = 1\text{-}d(i, j)/d_{max}$$

where d(i, j) is the Eluclidean Distance between color/edge bins i and j, and dmax is the maximum distance.

Then the similarity value S_M is calculated as follows:

$$S_M(X,Y)=Z^t A Z$$

where matrix Z = [z(k)] is the bin by bin difference between image X and Y and

$$z(k) = Xk-Yk, k=1, 2,....., N$$

Z^t is the transpose of Z.

4. Result and Discussion:

In this section an integrated approach combing color, HSV features and symmetry analysis for image retrieval has been designed and implemented. Various experiments have been carried out to evaluate the performance of this integrated approach. An image database of 654 JPEG format image is taken with different types of object in different image size. Results are obtained by taking a query image is given to the CBIR system and as an output we get images from image database with minimum distance with the query image.

Some of the results are shown below first of all the query image is shown then 8 image similar to query image returned by CBIR calculation are shown. After the displaying of retrieved image how many images are perceptually similar out of 8 images is discussed and collective results for all the experiments are tabulated at the last. Image 'glasgow3.jpg' is shown in Fig 1.This is a coloured image and its RGB planes are shown in Fig 2 (a) and there respective histograms are shown in Fig 2(b).

Fig. 1. Image of Glasgow.jpg

Fig 2(a) Gray level image of 'glasgow3.jpg' showing intensity in RGB planes (from left to right).

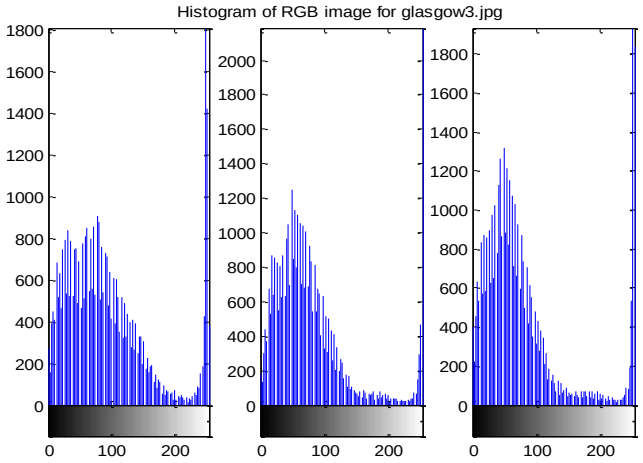

Fig. 2(b) Image histograms of images of RGB plane shown in Fig 2(a) for 100 bins.

HSV image of glasgow3.jpg

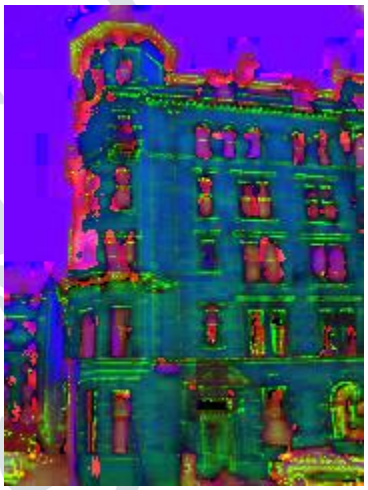

Fig. 3(a). HSV image based on hue saturation and value of RGB image 'glasgow3.jpg'

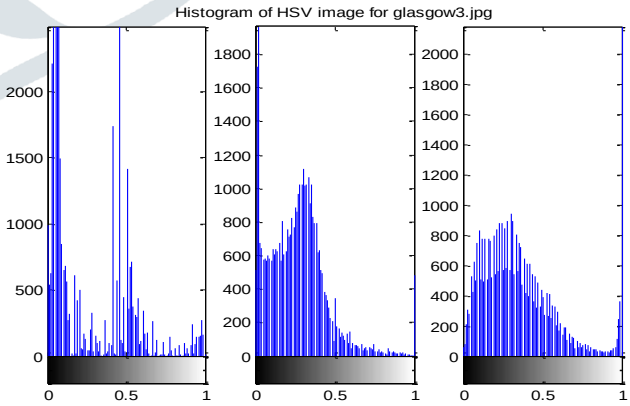

Fig. 3(b). Image histograms of images of HSV plane shown in Fig 3(a) for 100 bins.

Fig 2(b) indicates the intensity distribution of RGB planes in form of histogram. The histogram preserves the information in the form of no. of pixels for each image color contents. Hence histogram can be utilized to compare the feature of images. Fig 3(a) shows the transformation of same is but in HSV plane and Fig 3(b) indicates the histograms of HSV image.

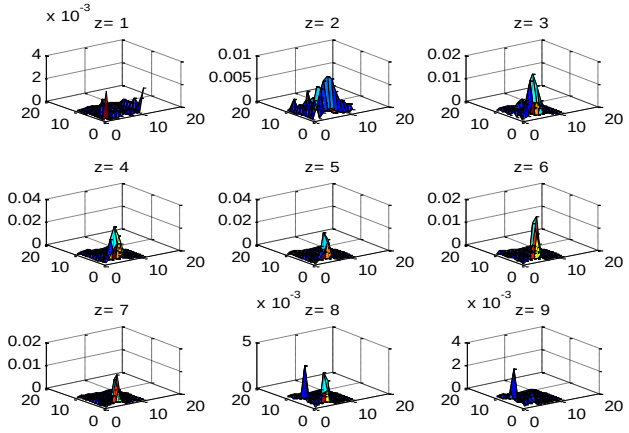

Fig 4. 3D (HSV) histogram of the query image for image of Fig 1.

3D histogram of HSV transformed image showing distribution of pixels in respect to their 11x11x11 bin distribution. Similarly we will show the 2D and 3d histograms of another image to demonstrate that the pattern of histogram pixel density distribution varies from image to image.

5. Conclusion:

In this work, image retrieval methods based on color, shape and spatial analysis are investigated. We have designed and implemented a prototype to retrieve a particular image from an image database. We have designed an indexing methods based on different criteria. We introduce an integrated method that calculates the similarity value between two images. We then evaluate the performance and compare the characteristic of each image retrieval approach.

References:

[1] Wang, j. z., boujemaa, n., del bimbo, a., geman, d., hauptmann, a., and tesic, J. 2006. Diversity in multimedia information retrieval research. In Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR) at the International Conference on Multimedia.

[2] Mokhtarian, F. 1995. Silhouette-Based isolated object recognition through curvature scale space. IEEE Trans. Pattern Anal. Mach. Intell. 17, 5, 539–544.

[3] Petrakis, e. and faloutsos, A. 1997. Similarity searching in medical image databases. IEEE Trans. Knowl. Data Eng. 9, 3, 435–447.

[4] Smith, j. and chang, S.-F. 1997a. Integrated spatial and feature image query. IEEE Trans. Knowl. Data Eng. 9, 3, 435–447.

[5] Chang, s., shi, q., and yan, c. 1987. Iconic indexing by 2-D strings. IEEE Trans. Pattern Anal. Mach. Intell. 9, 3, 413–427.

[6] Arnold W.M. Smeulders, M. Worring, S. Santini, Ramesh Jain, A. Gupta, "Content-Based Image Retrieval at the end of the Early years," IEEE transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1349-1380, 2000.

[7] P. Kruizinga, N. Petkov and S.E. Grigorescu, "Comparison of texture features based on Gabor filters," Proceedings of the 10th International Conference on Image Analysis and Processing, Venice, Italy, pp.142-147,1999.

[8] B.S. Manjunath and W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 18, No. 8, pp. 837-842, Aug, 1996

[9] SMEULDERS, A. W.,WORRING, M., SANTINI, S., GUPTA, A., , AND JAIN, R. 2000. Content based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22, 12, 1349–1380.

[10] Aigrain, p., zhang, h., and petkovic, d. 1996. Content-based representation and retrieval of visual media: A review of the state-of-the-art. Multimed. Tools Appl. 3, 3, 179–202.

[11] Rui, y., huang, t., and chang, S.-F. 1999. Image retrieval: Current techniques, promising directions and open issues. J. Visual Commun. Image Represent. 10, 1, 39–62.

[12] Snoek, c. g. m. and worring, M. 2005. Multimodal video indexing: A review of the state-of-the-art. Multimed. Tools Appl. 25, 1, 5–35.

[13] Zhou, x. s. and huang, T. S. 2003. Relevance feedback in image retrieval: A comprehensive review. Multimed. Syst. 8, 536–544