International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Image Mining for Retrieval of Image using Histogram Technique

Arun JB#* and Reshu Choudhary^

#Govt, Polytechnic College Ajmer, India
^Bhagwant University Ajmer, India

Received 20 Sept 2017, Accepted 21 Nov 2017, Available online 22 Nov 2017, **Vol.7, No.6 (Nov/Dec 2017)**

Abstract

Data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Image mining is simply an expansion of data mining in the field of image processing. Image mining handles with the hidden knowledge extraction, image data association and additional patterns which are not clearly accumulated in the images. Retrieval of required-query-similar images from accessible digital images is done by proposed improved histogram method. This method gave better precision value of the retrieved resultant images with 32% improvement. It is designed and implemented on MATLAB and is tested with the images of various databases.

Keywords: *Image mining, data mining, histogram technique, image matching and image retrieval*

1. Introduction

The success of the digital revolution and the growth of the internet have ensured that huge volumes of high-dimensional multimedia data are available all around us. This information is often mixed, involving different data types such as text, image, audio, speech, hypertext, graphics, and video components interspersed with each other. The World Wide Web has played an important role in making the data, even from geographically distant locations, easily accessible to users all over the world. However, often most of this data are not of much interest to most of the users. The problem is to mine useful information or patterns from the huge datasets. Data mining refers to this process of extracting knowledge that is of interest to the user. Data mining is an evolving and growing area of research and development, both in academia as well as in industry. It involves interdisciplinary research and development encompassing diverse domains.

The first generation of data mining algorithms has been demonstrated to be of significant value across a variety of real-world applications. But these work best for problems involving a large set of data collected into a single database, where the data are described by numeric or symbolic features. Here the data invariably do not contain text and image features interleaved with these features, and they are carefully and cleanly collected with a particular decision-making task in mind (J. Han and M. Kamber, 2001). Development of new generation algorithms is expected to encompass more diverse sources and types of data that will support mixed-initiative data mining, where human experts collaborate with the computer to form hypotheses and test them.

Image Mining

Traditional data mining techniques have been developed mainly for structured data types. The image data type does not belong to this structured category, suitable for interpretation by a machine, and hence the mining of image data is a challenging problem. Content of an image is visual in nature and the interpretation of the information conveyed by an image is mainly subjective, based on the human visual system. Image data have been used for machine vision, based on extraction of desired features from an image and interpretation of these features for particular applications. The research and development of mining image data is relatively new, and has become an emerging field of study today. Most of the activities in mining image data have been in the search and retrieval of images based on the analysis of similarity of a query image or its features with the entries in the image database (S. Mitra *et al*, 2002). The image retrieval systems can be broadly categorized into two categories based on the type of searches, using either description of an image or its visual content. In the first category, the images are described based on user-defined texts. The images are indexed and retrieved based on these rudimentary descriptions, such as their size, type, date and time of capture, identity of owner, keywords, or some text description of the image. The text-based descriptions of the images are usually typed

*Corresponding author **Reshu Choudhary** (ORCID ID: 0000-0002-5377-1242) is Research Scholar and **Arun JB** is working a Lecturer

manually for each image by human operators, because the automatic generation of keywords for the images is difficult without incorporation of visual information and feature extraction. As a result, this is a very labor-intensive process and is impractical in today's multimedia information age. Moreover, since the descriptions of images are very much subjective, the automated process to generate a text based description for indexing of the images could be very inaccurate and incomplete. The second category of similarity based image retrieval process is called Content Based Image Retrieval (CBIR). In CBIR systems, the images are searched and retrieved based on the visual content of the images. Based on these visual contents, desirable images features can be extracted and used as index or basis of search. Content-based image retrieval is highly desirable and has increasingly become a growing area of study towards the successful development of image mining techniques.

Content-Based Image Retrieval

There are, in general, three fundamental modules in a content-based image retrieval system. These are:

- Visual content or feature extraction,
- Multidimensional indexing, and
- Retrieval.

The images in an image database are indexed-based on extracted inherent visual contents (or features) such as color, texture, pattern, image topology, shape of objects and their layouts and locations within the image, etc. An image can be represented by a multidimensional vector of the extracted features from the image (B.S. Manjunathi, and W.Y. Ma, 1996). The feature vector actually acts as the signature of the image. This feature vector can be assumed to be associated to a point in the multidimensional space. The query image can be analyzed to extract the visual features and can be compared to find matches with the indices of the images stored in the database. The extracted image features are stored as meta-data, and images are indexed based on these meta-data information. This meta-data information comprises some measures of the extracted image features (M. Stricker and M. Orengo, 1995).

The CBIR systems architecture is essentially divided into two parts. In the first part, the images from the image database are processed off-line. The features from each image in the image database are extracted to form the meta-data information of the image, in order to describe the image using its visual content features (T. Ojala *et al*, 1996). Next these features are used to index the image, and they are stored into the meta-data database along with the images. In the second part, the retrieval process is depicted. The query image is analyzed to extract the visual features, and these features are used to retrieve the similar images from the image database. Rather than directly comparing two images, similarity of the visual features of the query image is measured with the features of each image stored in the meta-data database as their signatures. Often the similarities of two images are measured by computing the distance between the feature vectors of the two images. The retrieval system returns the first 'k' images, whose distance from the query image is below some defined threshold. Several image features have been used to index images for content-based image retrieval systems (A.K. Jain and R.C. Dubes, 1998). Most popular amongst them are color, texture, shape, image topology, color layout, region of interest, etc.

2. Literature Survey

The digital revolution has made digitized information easy to capture, process, store, distribute, and transmit. With significant progress in computing and related technologies and their ever-expanding usage in different walks of life, huge amount of data of diverse characteristics continue to be collected and stored in databases. Data mining is an attempt to make sense of the information explosion embedded in this huge volume of data. The advanced database management technology of today is enabled to integrate different types of data, such as image, video, text, and other numeric as well as non-numeric data, in a provably single database in order to facilitate multimedia processing. As a result, traditional adhoc mixtures of statistical techniques and data management tools are no longer adequate for analyzing this vast collection of mixed data.

The current Internet technology and its growing demand necessitates the development of more advanced data mining technologies to interpret the information and knowledge from the data distributed all over the world. In the 21st century this demand will continue to grow, and the access of large volumes of multimedia data will become a major transforming theme in the global society (A. Pentland *et al*, 1996).

The rapid growth of interest in data mining is due to the:

(i) Advancement of the Internet technology and wide interest in multimedia applications in this domain.
(ii) Falling cost of large storage devices and increasing ease of collecting data over networks.
(iii) Sharing and distribution of data over the network, along with adding of new data in existing data repository.
(iv) Development of robust and efficient machine learning algorithms to process this data.
(v) Advancement of computer architecture and falling cost of computational power, enabling use of computationally intensive methods for data analysis.
(vi) Inadequate scaling of conventional querying or analysis methods, prompting the need for new ways of interaction.
(vii) Strong competitive pressures from available commercial products, etc.

A. Image Mining

Image mining deals with the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in the images (A. Pentland *et al*, 1996). It is more than just an extension of data mining to the image domain. Image mining is an interdisciplinary endeavor that draws upon expertise in computer vision, pattern recognition, image processing, image retrieval, data mining, machine learning, database, artificial intelligence, and possibly compression.

Intelligently classifying an image by its content is an important way to mine valuable information from a large image collection. Since the storage and communication bandwidth required for image data is pervasive, there has been a great deal of activity in the international standard committees to develop standards for image compression. It is not practical to store the digital images in uncompressed or raw data form (J. Kacprzyk and S. Zadrozny, 1998). Image compression standards aid in the seamless distribution and retrieval of compressed images from an image repository.

Recent increase in the size of multimedia information repositories, consisting of mixed media data, has made CBIR an active research area. Unlike traditional database techniques which retrieve images based on exact matching of keywords, CBIR systems represent the information content of an image by visual features such as color, texture, and shape, and they retrieve images based on similarity of features (S. Russell and W. Lodwick, 1999). Frigui has developed an interactive and iterative image retrieval system that takes into account the subjectivity of human perception of visual content (B. Turksen, 1998). The smooth transition in the user's feedback is modeled by continuous fuzzy membership functions. Medasani and Krishnapuram have designed a fuzzy approach to handle complex linguistic queries consisting of multiple attributes.

B. Image Retrieval

Image mining requires that images be retrieved according to some requirement specifications. The requirement specifications can be classified into three levels of increasing complexity (Y. Bengio *et al,* 2000):
Level 1: comprises image retrieval by primitive features such as color, texture, shape or the spatial location of image elements.
Level 2: comprises image retrieval by derived or logical features like objects of a given type or individual objects or persons.
Level 3: comprises image retrieval by abstract attributes, involving a significant amount of high level reasoning about the meaning or purpose of the objects or scenes depicted.
Rick Kazman and John Kominek propose three query schemas for image retrieval: Query by Associate Attributes, Query by Description, and Query by Image Content.

Commercially, IBM's QBIC system is probably the best known of all image content retrieval systems (R. Kohavi *et al*, 2002). It offers retrieval by any combination of color, texture or shape, as well as text keyword. More efficient indexing techniques, an improved user interface, and the ability to search grey-level images have been incorporated in the latest version. Virage is another well-known commercial system (S. Mitra *et al*, 2002). This is available as a series of independent modules, which system developers can build into their own programs. Excalibur, by virtue of its company's pattern recognition technology, offers a variety of image indexing and matching techniques (Y. Bengio *et al*, 2000).

C. Image Indexing

While focusing on the information needs at various levels, it is also important to provide support for the retrieval of image data with a fast and efficient indexing scheme. Typically, the image database to be searched is large and the feature vectors of images are of high dimension, search complexity is high. Two main approaches are: reducing dimensionality or indexing high dimensional data. Reducing the dimensions can be accomplished using two well-known methods: the Singular Value Decomposition (SVD) update algorithm and clustering. The latter realizes dimension reduction by grouping similar feature dimensions together. High-dimensional indexing schemes includes SR-tree, TV-tree, X-tree and iMinMax (Y.C. Hu *et al*, 2003).

Current image systems retrieve images based on similarity. However, Euclidean measures may not effectively simulate human perception for certain visual content. Other similarity measures such as histogram intersection, cosine, correlation, etc., need to be utilized. One promising approach is to first perform dimension reduction and then use appropriate multi-dimensional indexing techniques that support Non-Euclidean similarity measures (V. Enireddy and K. K. Reddi, 2012). Develop an image retrieval system on Oracle platform using multi-level filters indexing (T.Y. Gajjar and N.C. Chauhan, 2012). The filters operate on an approximation of the high dimension data that represents the images, and reduces the search space so that the computationally expensive comparison is necessary for only a small subset of the data. Present a new compressed image indexing technique by using compressed image features as multiple keys to retrieve images (R. Krishnapuram *et al*, 2001).

The biggest issue for CBIR system is to incorporate versatile techniques so as to process images of diversified characteristics and categories. Many techniques for processing of low level cues are distinguished by the characteristics of domain-images. The performance of these techniques is challenged by various factors like image resolution, intra-image illumination variations, non-homogeneity of intra-region and inter-region textures, multiple and occluded objects etc. The other major difficulty, described as

semantic-gap in the literature, is a gap between inferred understanding / semantics by pixel domain processing using low level cues and human perceptions of visual cues of given image. In other words, there exists a gap between mapping of extracted features and human perceived semantics. The dimensionality of the difficulty becomes adverse because of subjectivity in the visually perceived semantics, making image content description a subjective phenomenon of human perception, characterized by human psychology, emotions, and imaginations. The image retrieval system comprises of multiple inter-dependent tasks performed by various phases. Inter-tuning of all these phases of the retrieval system is inevitable for over all good results. The diversity in the images and semantic-gap generally enforce parameter tuning & threshold-value specification suiting to the requirements. For development of a real time CBIR system, feature processing time and query response time should be optimized. A better performance can be achieved if feature-dimensionality and space complexity of the algorithms are optimized. Specific issues, pertaining to application domains are to be addressed for meeting application-specific requirements. Choice of techniques, parameters and threshold-values are many a times application domain specific e.g. a set of techniques and parameters producing good results on an image database of natural images may not produce equally good results for medical or microbiological images.

D. Color Histogram

The color histogram for an image is constructed by counting the number of pixels of each color. Each pixel is associated to a specific histogram only on the basis of its own color, and color similarity across different color dissimilarity in the same is not taken into account. Since any pixel in the image can be described by three components in a certain color space histogram, i.e., the distribution of the number of pixels for each quantized color, can be defined for each component. By default the maximum number of colors one can obtain using the histogram function is 256. Conventional color histogram (CCH) of an image indicates the frequency of occurrence of every color in an image. The appealing aspect of the CCH is its simplicity and ease of computation.

In 1999, by Ordonez & Omiecinski Rule mining has been implemented to huge image databases. There are two most significant techniques. The first technique is to mine from huge amount of images alone and the second technique is to mine from the integrated collections of images and related alphanumeric data. Image mining handles with all features of huge image databases which comprises of indexing methods, image storages, and image retrieval, all regarding in an image mining system by Missaoui & Palenichka in 2005. Mining results are obtained after matching the model description with its complementary symbolic description. The symbolic description might be just a feature or a set of features, a verbal description or phrase in order to identify a particular semantic in 2007 by Fernandez. Developments in area of image acquisition and storage technique have shown the way for incredible growth in extensively large and detailed image databases. The images which are available in these databases, if examined, can provide valuable information to the human users. Image mining facilitates the extraction of hidden information, image data association, or other patterns not clearly accumulated in the images. Image mining is an interdisciplinary effort that provides significant application in the domain of machine learning, image processing, image retrieval, data mining, database, computer vision, and artificial intelligence (A. Kannan *et al*, 2010; D.S. Zhang and G. Lu, 2000). Even though the growth of several applications and techniques in the individual research domain mentioned above, research in image mining has to be explored in investigated the research problems in image mining, modern growth in image mining, predominantly, image mining frameworks, modern techniques and systems. Common pattern identical, pattern identification and data mining models with the intention that a real life scene/image can be associated to a particular category, assisting in different prediction and forecasting mechanisms. It is a three-step procedure i.e. image gathering, learning and classification. Image mining approach using clustering and data compression techniques was projected by Pattnaik in 2008. Satellite images of clouds play a substantial role in forecasting weather conditions (Y. Chen *et al*, 2005). Frequency of image acquirement ranges from one image per minute to another image per hour based on the climatic environment. Decision tree based image processing and image mining technique was projected by Kun-Che in 2009. Important information can be hidden in images, conversely, few research talks about data mining on them.

3. Proposed Method

The Color histogram information is preceded for retrieval of images in this proposed method. For Computing an Image color histogram, the different color axes are divided into a number of bins. A three dimensional 256x256x3 RGB histogram would therefore contain a total of 196608 such bins are converted into two dimensional histogram. When indexing the image, the color of each pixel is found, and the corresponding bin's count is incremented by one.

In order to compare histograms of database image and query image, first need to generate specific codes for all histogram bins. In this method, (r: 0-255, g: 0-255, b: 0-255) codes were generated for RGB histogram bins. When the images have been quantized into histograms, a method of comparing these is needed. Two color histograms are divided into bins. The distance of each bin is calculated and square of it is used to calculate overall distance D and take same for between every combination of bins.

This approach is applied over all three color histograms as give in following equations:

$$D_r = \sum_{i=1}^{256} \left(H_{r1}(i) - H_{r2}(i) \right)^2 \quad (1)$$

$$D_g = \sum_{i=1}^{256} \left(H_{g1}(i) - H_{g2}(i) \right)^2 \quad (2)$$

$$D_b = \sum_{i=1}^{256} \left(H_{b1}(i) - H_{b2}(i) \right)^2 \quad (3)$$

Where: D_r, D_g and D_b are the calculated distance for red, green and blue colors, H_{r1}, H_{g1} and H_{b1} are the histogram bins for red green and blue colors of query image, H_{r2}, H_{g2} and H_{b2} are the histogram bins for red, green and blue colors of database image respectively.

After obtaining the Distances D_r, D_g and D_b, these distances are compared with threshold distances and if all the computed values are less than the threshold values than images are matched and subsequently results are produced. The threshold condition is $D_r < D_{rth}$, $D_g < D_{gth}$ and $D_b < D_{bth}$. Where the D_{rth}, D_{gth} and D_{bth} are the threshold values for red, green and blue colors.

The performance of retrieval system can be measured in terms of its recall and precision. There are many methods exist for comparing histograms with different precision and recall value. Here c we our histogram comparison method with the result present satisfactory precision and recall values (Arun JB and Reshu Choudhary, 2013). Recall measures the ability of the system to retrieve all the models that are relevant, while precision measures the ability of the system to retrieve only the models that are relevant. It has been reported that the histogram gives the best performance through recall and precision value. They are defined as:

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \quad (4)$$

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images}} \quad (5)$$

4. Experimental Results

The proposed method is tested on database of 158 images. Fig.1 shows the precision and recall values for different thresholds and find the value of crossover point is 0.66. The results show 100% precision and recall value.

The Result is 32% improved as when we use only the sum of distances method as shown in Fig 2 (Arun JB and Reshu Choudhary, 2013). In that case precision and recall crossover value is 0.50 and gave only 50% precision, hence in our proposed method precision rate is significantly improved.

For retrieving similar images, the value of threshold is set according the values of histogram bins. If a selected threshold values satisfied the earlier discussed conditions then search results of the images are shown in Fig 3.

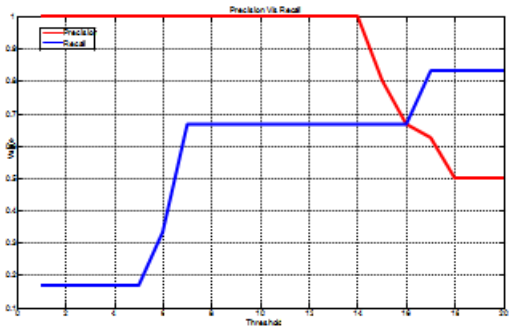

Fig 1. Plot for precision and recall values of proposed method

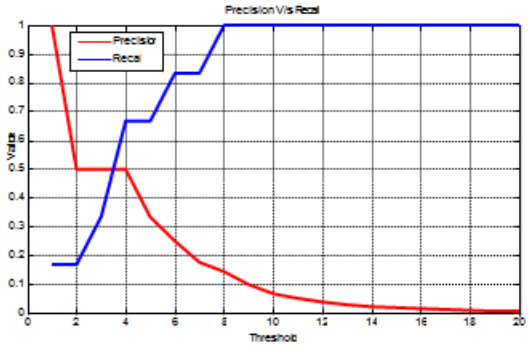

Fig 2. Precision and recall plot (Arun JB and Reshu Choudhary, 2013)

Fig 3. Retrieved Results of proposed method

This proposed method gave 87.5% accuracy of precision from database images.

Conclusion

Image mining presents special characteristics due to the richness of the data that an image can show. Effective evaluation of the results of image mining by content requires that the user point of view is used on the performance parameters. The query image is compared to each of database images to determine whether they are equivalent or not by comparing with all features. Some of the proposed techniques produced good results and some may not. This paper compared proposed technique with improved histogram technique. When only color histogram is

considered as retrieval parameter in CBIR gives 32% improved average retrieval precision and recall crossover values. Improvement in retrieval efficiency and improvement of speed by applying other algorithms can be addressed in future works.

References

Han, J. and Kamber, M. (2001), Data Mining: Concepts and Techniques, San Diego: Academic Press.

Mitra, S., Pal, S. K. and Mitra, P. (2002), Data Mining in Soft Computing Framework:A Survey, IEEE Transactions on Neural Networks, vol. 13, pp. 3-14.

Manjunathi, B.S. and Ma, W.Y. (1996), Texture Features for Browsing and Retrieval of Image Data, in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 8.

Stricker, M. and Orengo, M. (1995), Similarity of Color Images, in Proc. SPIE Storage and Retrieval for Image and Video Databases, pp. 381-392.

Ojala, T., Pietikainen, M. and Harwood, D. (1996), A Comparative Study of Texture Measures with Classification Based on Feature Distribution, Pattern Recognition, pp. 51-59.

Jain, A.K. and Dubes, R.C. (1998), Algorithms for Clustering Data, Prentices Hall Advanced Reference Series.

Pentland, A., Picard, R.W. and Sclaro, S. (1996), Photobook: Content-Based Manipulation of Image Databases in Multimedia Tools and Applications, editor Borko Furht, Kluwer Academic Publishers, Boston, pp. 43-80.

Kacprzyk, J. and Zadrozny, S. (1998), Data Mining via Linguistic Summaries of Data: An Interactive Approach, in Proceedings of IIZUKA 98 (Fukuoka, Japan), pp. 668-671.

Russell, S. and Lodwick, W. (1999), Fuzzy Clustering in Data Mining for Telco Database Marketing Campaigns, in Proceedings of NAFIPS 99 (New York),pp. 720-726.

Turksen, B. (1998), Fuzzy Data Mining and Expert System Development, in Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (San Diego, CA), pp. 2057-2061.

Bengio, Y., Buhmann, J. M., Embrechts, M. and Zurada, J. M. (2000), Introduction to the Special Issue on Neural Networks for Data Mining and Knowledge Discovery, IEEE Transactions on Neural Networks, vol. 11, pp. 545-549.

Kohavi, R., Masand, B., Spilipoulou, M. and Srivastava, J. (2002), Web Mining, Data Mining and Knowledge Discovery, vol. 6, pp. 5-8.

Hu, Y.C., Chen, R.S. and Tzeng, G.H. (2003), Finding Fuzzy Classification Rules using Data Mining Techniques, Pattern Recognition Letters, vol. 24, pp. 509-519.

Enireddy, V. and Reddi, K. K. (2012), A Data Mining Approach for Compressed Medical Image Retrieval. International Journal of Computer Applications (0975 – 887) Volume 52– No.5.

Gajjar, T.Y. and Chauhan, N.C. (2012), A Review on Image Mining Frameworks and Techniques, International journal of computer science and information technologies,Vol 3.

Krishnapuram, R., Joshi, A., Nasraoui, O. and Yi, L. (2001), Low Complexity Fuzzy Relational Clustering Algorithms for Web Mining, IEEE Transactions on Fuzzy Systems, vol. 9, pp. 595-607.

Kannan, A., Mohan, V. and Anbazhagan, N. (2010) ,Image Clustering and Retrieval using Image Mining Techniques, IEEE International Conference on Computational Intelligence and Computing Research.

Zhang, D.S. and Lu, G. (2000), Content Based Image Retrieval Using Texture Features, In Proc. Of First IEEE Pacific-Rim Conference on Multimedia PCM, pp.392-395.

Chen, Y., Wang, J.Z. and Krovetz, R. (2005), Cluster Based Retrieval of Images by Unsupervised Learning. IEEE Transaction on Image Processing, Vol 14, pp.1187-1199.

JB, A. and Choudhary, R. (2013), Image Retrieval using Color Histogram Matching Technique, IJTAR,ISSN:2249-8141,Vol.3, No. 3.