IJCSMS International Journal of Computer Science & Management Studies, Special Issue of Vol. 12, June 2012
ISSN (Online): 2231 –5268
www.ijcsms.com

Content Based Image Retrieval through Clustering

Sandhya[1], Preeti Gulia[2]

[1]M.tech Student, Department of Computer Science and Applications,
M. D. University, Rohtak-124001, Haryana, India
sandhyaphogat@gmail.com

[2]Assistant Professor, Department of Computer Science and Applications,
M. D. University, Rohtak-124001, Haryana, India

Abstract

Content-based image retrieval (CBIR) is a technique used for extracting similar images from an image database. CBIR system is required to access images effectively and efficiently using information contained in image databases. Here, K-Means is to be used for Image retrieval. The K-means method can be applied only in those cases when the mean of a cluster is defined. The K-means method is not suitable for discovering clusters with non-convex shapes or clusters of very different size. In this paper, CBIR, clustering and K-Means are defined. With the help of these, the data consisting images can be grouped and retrieved.

Keywords: CBIR, Image Retrieval, Clustering.

1. Introduction

Applications involving the search and management of digital images have increased tremendously over the last few years. There has been an explosive growth in the acquisition and use of images in health care data. Interest in image retrieval has increased in large part due the rapid growth of the World Wide Web. Image databases exist for storing art collections, satellite images, medical images and many other real-time applications. Image databases can be so large, which contain hundreds of thousands of images. However, it is not possible to access or make use of this information unless it is organized to allow for efficient browsing and retrieval. Content-based image retrieval (CBIR) is a technique used for extracting similar images from an image database. Color, texture and shape features have been used for describing image content. CBIR system is required to effectively and efficiently access images using information contained in image databases. A CBIR system uses information from the content of images for retrieval and helps the user to retrieve database images relevant to the contents of a query image.

2. Related Work

P. S. Hiremath et al. [6] specify that there are some commercial image search engines available on the Web such as Google Image Search and AltaVista Image Search. Most of them employ only the keyword based search and hence the retrieval result is not satisfactory. With the advances in image processing, information retrieval, and database management, there have been extensive studies on content-based image retrieval (CBIR) for large image databases.

Chen, Y. and Wang, J. Z. et al. [8] provide systems retrieve images based on their visual contents. Earlier efforts in CBIR research have been focused on effective feature representations for images. The visual features of images, such as color, texture, and shape features have been extensively explored to represent and index image contents, resulting in a collection of research prototypes and commercial systems. There are also some integrated search engines employing both the keyword-based search and content-based image.

Maria Halkidi et al.[4] define the clustering as the most important *unsupervised learning* problem. It deals with finding a *structure* in a collection of unlabeled data. The basic definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. It is nothing but the grouping of objects where the objects in a cluster are similar to each other and dissimilar to objects in other clusters. Clustering is usually based on a distance metric that allows us to assess distances between objects and distances between clusters. The distance metrics used are the Euclidean and Kullback–Leibler.

Jiawei Han et.at [1] defines the hierarchical clustering techniques are based on the use of a proximity matrix indicating the similarity between every pair of data points to be clustered. The end result is a tree of clusters, called a dendrogram representing the nested grouping of patterns and similarity levels at which groupings change. It proceeds successively by either merging smaller clusters into larger ones (agglomerative, bottom-up), or by splitting larger clusters (divisive, top-down). By cutting the dendrogram at a desired level, a clustering of data items into disjoint groups is obtained. The clustering methods differ in regards to the rules by which two small clusters are merged or a large cluster is split. Some of the hierarchical algorithms include Cobweb, Cure and Chameleon etc.

3. Content Based Image Retrieval (CBIR)

Content-based image retrieval (CBIR) is an image retrieval system, which aims at avoiding the use of textual descriptions and instead retrieves images based on their visual similarity to a user supplied query image or user-specified image features. Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision to the image retrieval problem, that is, the problem of searching for digital images in large databases. "Content-based" means that the search will analyze the actual contents of the image. The term 'content' in this context might refer colors, shapes, textures, or any other information that can be derived from the image itself. Without the ability to examine image content, searches must rely on metadata such as captions or keywords, which may be laborious or expensive to produce.

3.1 Clustering

In CBIR, clustering is also used. Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. The figure 3.1 demonstrates the clustering of balls of same colour. There are a total of 10 balls which are of three different colours. We are interested in clustering of balls of the three different colours into three different groups.

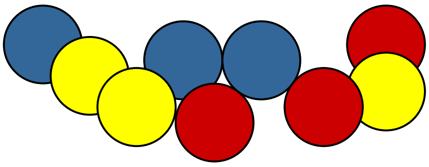

Figure 3.1: Objects before clustering

The balls of same colour are clustered into a group as shown in figure 3.2

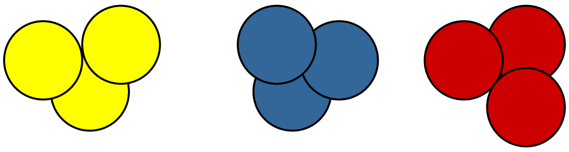

Figure 3.2: Objects after clustering

Thus, Clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.

1.3 K-Means

The most well known and commonly used partitioning method is K-means. The K-means algorithm takes the input parameter, K, and partitions a set of n object into K clusters so that the resulting intracluster similarity is high but the intercluster simililarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the clusters center of gravity.

The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence

- Determine the centroid coordinate
- Determine the distance of each object to the centroids
- Group the object based on minimum distance

Iterate until stable = no object move group:

4. Conclusion

The basic image retrieval can be done through K-Means. If user wants to search for images, image index have to be provided. The system will extract image feature for this query. It will compare these features with the images that are in database. Relevant results will be displayed to the user. The K-means method, however, can be applied only when the mean of a cluster is defined. This may not be the case in some applications, such as when data with categorical attributes are involved. The necessity for user to specific K, the number of clusters, in advance can be seen as a disadvantage. The K-means method is not suitable for discovering clusters with non-convex shapes or clusters of very different size.

The enhancements made to the casual image retrieval system can be called as CBIR-C system. CBIR-C system architecture is decomposed as fallows

- Data Collection phase
- K-Means clustering
- CBIR-C phase
- Input/output phase

From these CBIR-C system performance enhancement is improved, thus image retrieval can be done faster than available image retrieval systems.

The CBIR - C system is carried out for lesser amount of data and is also restricted to a particular domain. This system can further extend to any domain or may even be used to carry analysis on other diseases.

5. References

[1] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 2001.

[2] Arun K.Pujari, “Data Mining”, Universities Press (India) Ltd., 2001.

[3] Margaret H.Dunham, “Data Mining: Introductory and Advanced Topics”, Pearson education, 2003.

[4] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis, “On Clustering Validation Techniques”, A survey on KDD and clustering Techniques, Dept of Informatics, Athens University of Economics & Business.

[5] A.K. Jain. M.N. Murty, P.J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[6] P. S. Hiremath , Jagadeesh Pujari, “Content Based Image Retrieval using Color, Texture and Shape features”, 15th International Conference on Advanced Computing and Communications.

[7] Mei-Ling shyu, shu-Ching chen, Min Chen, Chengcui Zhang, “ A Unified Frame work for Image Database Clustering and Content-based Retrieval”, ACM Digital Library, MMDB, November 2004.

[8] Chen, Y. and Wang, J. Z. ,“A Region-based Fuzzy Feature Matching Approach to Content-based Image Retrieval”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, No. 9 , September 2002.

[9] Apostol Natsev, Rajeev Rastogi, and Kyuseok Shim, “WALRUS: A Similarity Retrieval Algorithm for Image Databases”, IEEE Transactions on Knowledge and Data Engineering, Vol.16, no.3, March 2004.

[10] Jia Wang, Wen-jam Yang* and Raj Acharya, “Color Clustering Techniques for Color-Content-Based Image Retrieval from Image Databases”, In Proceedings of IEEE International Conference on Multimedia and Expo(ICME’00), 1997, 114-121.

[11] Safar, M., Shahabi, C. and Sun, X. “Image Retrieval by Shape: A Comparative Study”, In Proceedings of IEEE International Conference on Multimedia and Expo (ICME’00), 2000, 141-144.

[12] Stehling, R. O., Nascimento, M. A., and Falcao, A. X. , “On Shapes of Colors for Content-based Image Retrieval”, In ACM International Workshop on Multimedia Information Retrieval (ACM MIR’00), 2000, 171-174.

[13] Zhang, D. S. and Lu, G, “Generic Fourier Descriptors for Shape-based Image Retrieval”, In Proceedings of IEEE International Conference on Multimedia and Expo (ICME’02), 1 (2002), 425-428.

[14] Shyu, M.-L., Chen, S.-C., Chen, M., and Zhang, C, “Affinity Relation Discovery in Image Database Clustering and Content-based Retrieval”, Acce for publication (short paper), ACM International Conference on Multimedia, October 10-16, 2004.

[15] Yong Rui, Thomas Huang and Shih-Fu Chang, “Image Retrieval: Current Techniques, Promising Directions, and Open Issues”, Published in the Journal of Visual Communication and Image Representation.

[16] Sangoh Jeong, “Histogram-Based Image Retrieval”, A Project Report.

[17] www.KDnuggets.com – Web site for Data Mining and Knowledge Discovery.

[18] www.Mathworks.com - Web site for MATLAB functions.