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

Implementation of CBIR Method and its Architecture

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

In present world, the usage of internet and World Wide Web is rapidly increasing and due to which information storage and retrieval from databases become a important task. In this information images plays a great role. Therefore, the applications involving the search and management of digital images have gained interest of many researchers. Content-based image retrieval (CBIR) is a technique used for extracting similar images from an image database. Colour, 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 retrieve database images relevant to the contents of a query image.

Keywords: *CBIR System Design, System Architecture, Interface Unit.*

1. Introduction

Image databases exist for storing art collections, satellite images, medical images and many other real-time applications. Image databases can be huge, containing hundreds of thousands of images. To search an image in this vast unorganized collection is a tough job. Therefore, there must be a technique to organize this image collection so that image browsing and retrieval become efficient and easy.

But the process of image retrieval is not as easy as query processing. In Image retrieval, Shape and Color of an object plays an important role.

The field of image retrieval has been an active research area for several decades and has gained attraction in recent years as a result, large collection of digital images are growing day by day in health care centers. We are drowning in medical data but starving for knowledge to take the predictive measure.

1.1CBIR

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). 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.

2. Related Work

At present, 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. CBIR **P. S. Hiremath** [6] **Chen, Y. and Wang, J. Z.** [8] **Shyu, M.-L., Chen, S.-C., Chen, M., and Zhang**[14] **Yong Rui, Thomas Huang and Shih-Fu Chang** [15] 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 **P. S. Hiremath** [6] **Jia Wang, Wen-jam Yang and Raj Acharya** [10], texture, and shape **P. S. Hiremath** [6] **Safar, M., Shahabi, C. and Sun** [11] **Stehling, R. O., Nascimento, M. A., and Falcao,A. X**. [12] **Zhang, D. S. and Lu, G** [13] 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

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

search engines employing both the keyword-based search and content-based image.

In data mining there exist many techniques such as Association Rule Mining, Classification and prediction, Cluster Analysis, Outlier Analysis and Evolution Analysis. Out of which clustering has been identified as the best technique for the CBIR-C system.

3. Proposed Work

In this study, K-means clustering algorithm is to be used to proposed new architecture. The high level architecture and requirements of the proposed system are given below.

3.1 Proposed CBIR- C System Architecture

In this architecture, CBIR-C system is used to recognize pattern through knowledge discovery in database for image database. As it is not an easy task like retrieving text data, content base image retrieval through clustering technique has been used for pattern reorganization.

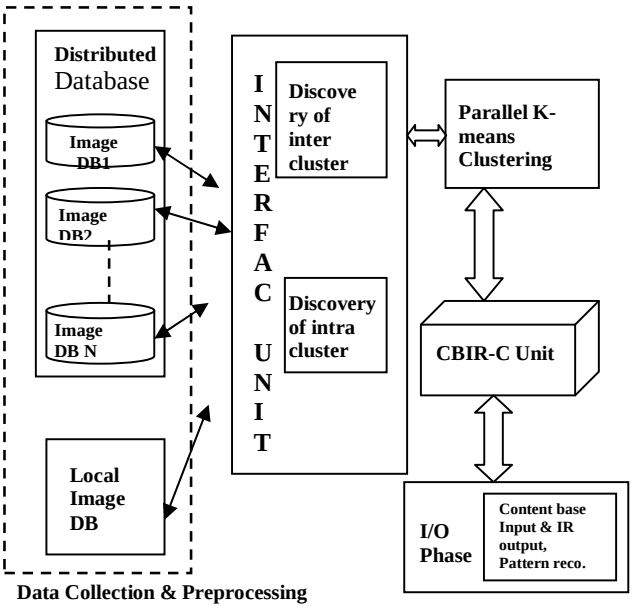


Fig: Architecture of the proposed CBIR-C system

The architecture of the proposed scheme is demonstrated in Figure. As can be seen from this figure, CBIR-C system framework contains major components, namely distributed database, Local database, interfacing unit, clustering process, CBIR-C unit, and Input Out put phase. In brief, given a set of image databases and the associated log data, a data mining process is conducted for intra-database knowledge discovery and summarization. Then the similarity measures among the databases are calculated via probabilistic reasoning. With the summarized database-level knowledge a conceptual database clustering process is carried out. Note that our framework is flexible in the sense that any database clustering strategy can be easily plugged in, as long as it has the capability to partition the databases into a set of database clusters. However, our conceptual database clustering process is highly effective. Thereafter, cluster-level knowledge summarization is applied to discover the intra-cluster knowledge and explore the inter-cluster relationships. Finally, image retrieval is conducted in the intra-cluster or inter-cluster level based on the obtained cluster-level knowledge. The description of proposed architecture is as follows:

3.1.1 Data Collection Phase

The image data is gathered from distributed environments such as Database, data warehouse, www, or other information repository: This is one or a set of databases, data warehouses, spreadsheets, or other kind of information repositories. Data cleaning [1] and data integration [1] techniques may be performed on the data. The database or data warehouse server is repository for fetching the relevant data, based on the user's data mining request.

The database management system used in the study was the Microsoft SQL Server 2000. This system was used for two reasons; the software used in analysis was compatible and efficient to use with the database management system, and the data to be analyzed was maintained in the database prior to the study.

Most of the image classification approaches apply clustering on a single database level. However, in a distributed environment, the number of image databases has increased enormously, and the query results may need to access several image databases at different locations. The principles of a distributed database system include executing a query as fast as possible or with as little cost as possible and allowing transparent location-independent accesses to the users in the applications. Hence, there is a strong need to analyze and discover summarized knowledge at the database level, i.e. database clustering. This phase is the combination of Distributed Database and Local Database.

Distributed Database: The image data distributed in World Wide Web are known as distributed database. These distributed databases are processed through content based image retrieval.

Local Database: The data bases which are stored locally from processing end are known as local

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

database.

3.1.2 Interface Unit

Our proposed framework performs cluster-based image retrieval adaptively in either the intra-cluster level or the inter-cluster level. Currently, many research efforts have been carried out to reduce the query search space via a clustering process. Once the clusters are obtained, the retrieval process is conducted within a certain cluster for a specific query, namely intra-cluster retrieval. However, as the principle of the clustering is to maximize the intra-cluster similarity and minimize the inter-cluster similarity by taking into account all the objects in the clusters (the so-called majority vote), it might not be an optimal solution for some specific objects. In other words, for an object O_i in cluster C_i its most related object(s) might not be included in C_i. In particular, this issue remains as a challenge for the cluster algorithm with a predefined cluster size (e.g., the single-link clustering method), where two related objects are partitioned into two clusters due to the limited cluster size or the predefined number of clusters. It is the combination of inter cluster and intra cluster image discovery.
Inter cluster: When the image data are clustered from World Wide Web it is known as discovery of inter cluster level.
Intra Cluster*:* When the image data are clustered from Local database it is known as discovery of Intra cluster level [7].

3.1.3 K-Means Clustering

The K- Means is known as a partition method as the user first predefines the number of clusters after which the algorithm partitions the data iteratively until a solution is found. As hierarchical clustering sorts data out into previously unknown clusters, K-Means actually assigns data between predefined partitions- the problem to solve is which cluster each data point belongs to. Thus the K-Means clustering is usually the most preferred method due to its simplicity.
K- Means algorithm is used for clustering based on the mean value of the records in the cluster. This algorithm assigns each point to the cluster whose center (also called the centroid) is nearest. The center is the average of all the points in the cluster- that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. This algorithm takes the input parameter k, and partitions the set of n objects into k clusters so that the resulting intra cluster similarity is high whereas the inter cluster similarity is low. Similarity is measured by the minimum distance between the points in the clusters, maximum distance between the points in the clusters and average distance between the points in clusters.
If the number of datasets is less than the number of clusters then each data is assigned as centroid of the cluster. Each centroid will have a cluster number. If the number of data is bigger than the number of cluster, for each data, the distance is calculated with all centroid and the minimum distance is found. This data is said to belong to the cluster that has the minimum distance from this data.AKD is proposed for the automatic detection of the k.

3.1.4 CBIR-C PHASE

Content-based image retrieval (CBIR) is a 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.

3.1.5 Input/output Phase

In this phase content based query is given through input unit and results obtained for that content based query are produced as output.

4. CBIR-C IMPLEMENTATION MODULES

Data collection module, CBIR-C module and K-means modules of the CBIR-C system architecture are discussed.

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

4.1 DATA COLLECTION MODULE

The image data is gathered from distributed environments such as Database, www, or other information repository: Data cleaning and data integration techniques may be performed on the data. This phase is the combination of Distributed Database and Local Database. The image data distributed in World Wide Web are known as distributed database. The data bases which are stored locally from processing end are known as local database. The data is composed of 18 different kinds of images such as brain, leg, hand, spinal card, tulip, satellite image, animal, airplane, flag, natural images etc.

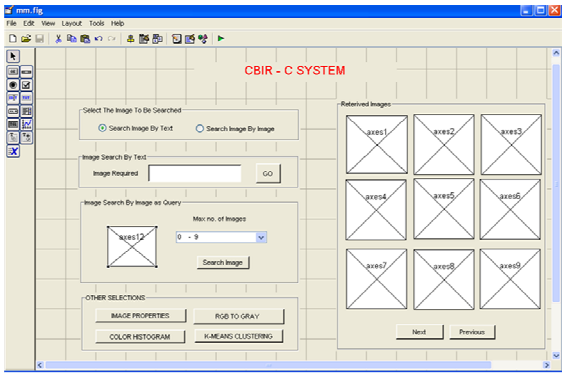
4.2 CBIR-C MODULES

CBIR system input module is designed in MATLAB as shown in the fig: SS:5. Retrieval test has been conducted on both medical images as well as natural images. The data is composed of 18 different kinds of images such as brain, leg, hand, spinal card, tulip, satellite image, animal, airplane, flag, natural images etc.
The system is developed as:

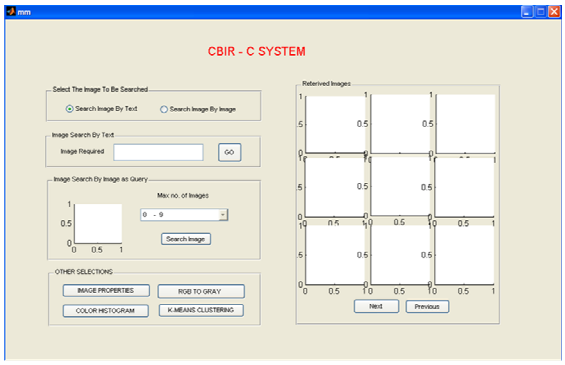
- User wants to search for, brain images
  - Submits an existing brain picture as query.
  - Submits his own sketch of brain as query.
- The system will extract image features for this query.
- It will compare these features with that of other images in a database.
- Relevant results will be displayed to the user.

The CBIR system GUI editor menu is shown in the SS: 5. The modules in the CBIR system are Image search based on the text (that is name of the image can be given for the search process) , giving image as query for image search, image properties, conversion of image property from RGB to gray color , color histogram and K-means.

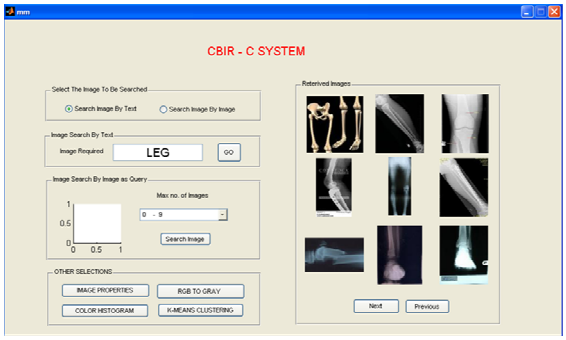
When ever user wants to search for a particular image that is present in the database, can be searched in two ways, search image based on text or image itself as query. As a first step in module one image search is based on the text. For this purpose image name is given as input to the input module and it is shown in the SS : 6. As input is given as image name such as brain, leg, spinal card, rose, tulip, sky etc the desired out put images are displayed . SS: 7 shows the out put for the desired leg image. As there are around 10000 images of different kind in the database, the module is developed such a manner that user can retrieve only some set of images ranges from 1-9, 10-19, 20-29. SS: 8 show the next step of the module that is next images in rang of 10-19. As a second step in module one, user can search image based on the image it self as query. For this purpose user gives image it self as query as input to the input module and it is shown in the SS: 9. Hear the brain image is given as input for the module and the desired out put for the query are displayed as shown in SS: 10.



SS 5: GUI editor menu for CBIR system

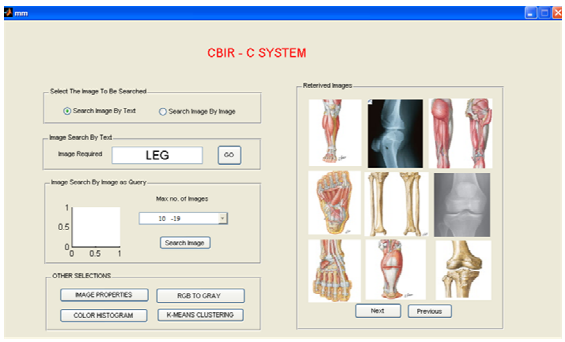


SS 6: Image search by text as input

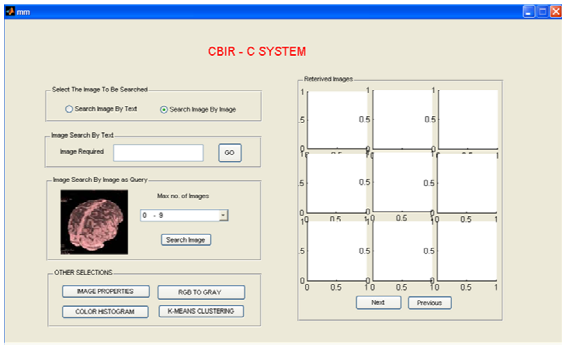


SS 7: sample out for the text (input is given as leg)

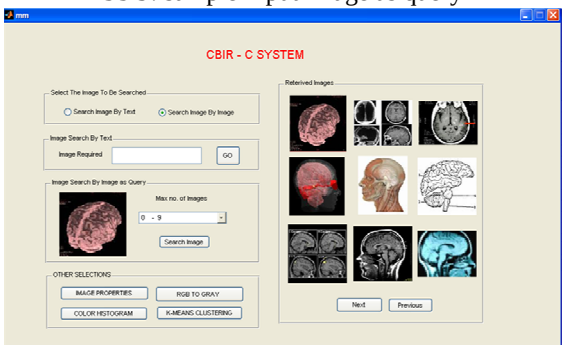
IJCSMS International Journal of Computer Science & Management Studies, Special Issue of Vol. 12, June 2012
ISSN (Online): 2231 –5268
www.ijcsms.com



SS 8: sample out for the text (input is given as leg)



SS 9: sample Input image as query



SS 10: sample out for the desired image brain

4.3 K-MEANS MODULES

The algorithm for K-Means is composed of the following steps:

Step - 1: Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step - 2: Assign each object to the group that has the closest centroid.

Step - 3: When all objects have been assigned, recalculate the positions of the K centroids.

Step - 4: Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Though K- Means clustering algorithm has got several advantages and is widely used in pattern recognition, it also has certain drawbacks. They are:

- It implies that the data clusters are ball-shaped because it performs clustering based on the Euclidean distance only
- There is the dead-unit problem. That is, if some units are initialized far away from the input data set in comparison with other units, they then immediately become dead without learning chance any more in the whole learning process.
- It needs to pre-determine the cluster number. When k equals to k', the k-means algorithm can correctly find out the clustering centres Otherwise, it will lead to an incorrect clustering result as depicted where some of seed points do not locate at the centres of the corresponding clusters. Instead, they are either at some boundary points among different clusters or biased from some cluster centres.

To overcome the drawback of k-means automatic Key detection is used.

The algorithm for AKD is composed of the following steps:

Step - 1: AKD scans through and identify unique objects from the database

Step - 2: Proceeds with general K-Means algorithm

Step - 3: If cluster is too large go to step 1

Step - 4: Proceeds with splitting concept

Step - 5: If cluster is two small go to step 1

Step - 6: Proceed with merging concept

Step - 7: Repeat until all the iteration are completed

5. ANALYSIS OF K GENERATION

The input and output modules of K-means are shown

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

in SS: 1 to SS: 4. The graph shown below in fig (a) is the graphical analysis of the implemented automatic K detection algorithm. The X plot represents the number of data points in each cluster and the Y plot represents the number of iterations carried out. The graph clearly shows the efficiency of the algorithm as the number of iteration rises. The cluster number K becomes constant after undergoing an initial merge and split procedure.

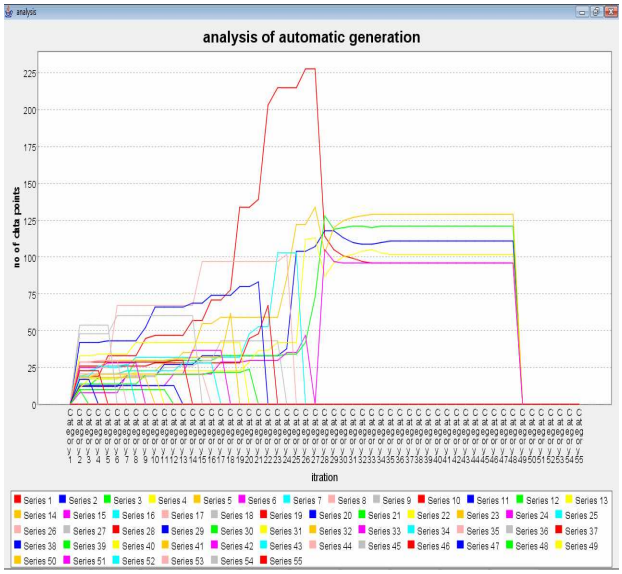


Fig :(a) Analysis of auto K algorithm

The alike medical records are grouped into similar clusters. The figs (b) and fig (c) represent the plot of different clusters. The X plot the attributes and the Y plot represents the data values of these attributes. The analysis of these clusters helps us identify meaningful patterns. For example already known features like the patients venous plasma glucose level is relative to the fasting sugar level can be obtained. Another recognized feature is that the patients who are chain smokers are prone to have more problems in diabetes than the normal patients.

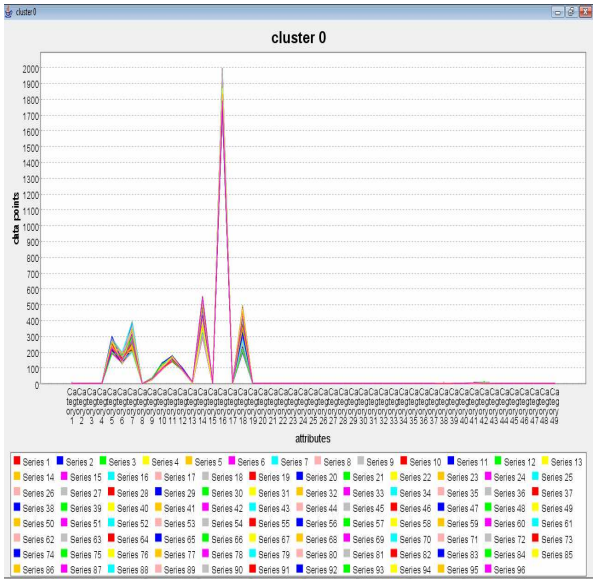


Fig: (b) Representation of cluster 0

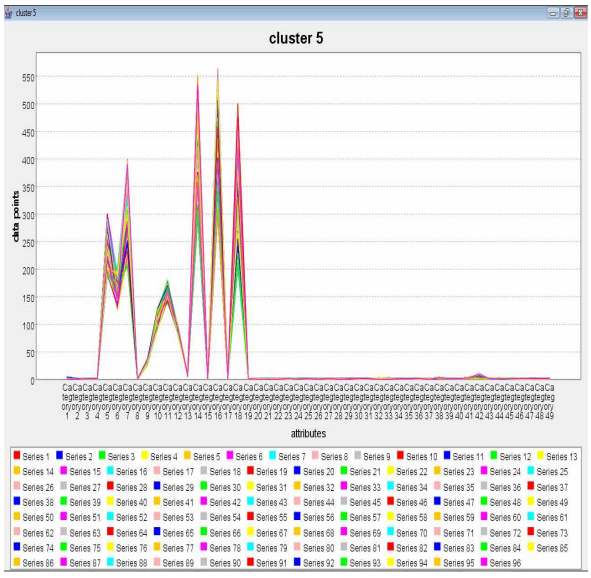


Fig :(c) Representation of cluster 5

6. 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

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

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.

7. 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.