International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
doi : https://doi.org/10.32628/CSEIT206385

# Study of Content Based Image Retrieval Using Data Mining Techniques

**Dr. S. Dhinakaran**

Associate Professor, Department of Computer Science, Rathinam College of Arts & Science, Coimbatore, Tamil Nadu, India

## ABSTRACT

The field of image retrieval has been an active research area for several decades and has been paid more and more attention in recent years as a result of the dramatic and fast increase in the volume of digital images. Content-based image retrieval (CBIR) is a new but widely adopted method for finding images from vast and un annotated image databases. In recent years, a variety of techniques have been developed to improve the performance of CBIR. In reaction to the needs of users, who feel problems connected with traditional methods of image searching and indexing, researchers focus their interest on techniques for retrieving images on the basis of automatically-derived features, often denoted as Content-Based Image Retrieval (CBIR). CBIR systems index the media documents using salient features extracted from the actual media rather than by textual annotations. Query by content is nowadays a very active research field, with many systems being developed by industrial and academic teams. Results performed by these teams are really promising. The situation gets diametrically different when we move our attention from the usual CBIR task, i.e. the retrieval of images which are similar (as a whole) to the query image, to the task "find all images that contain the query image". The proposed CBIR technique uses more than one clustering techniques to improve the performance of CBIR. This optimized method makes use of K-means and Hierarchical clustering technique to improve the execution time and performance of image retrieval systems in high dimensional sets. In this similarity measure is totally based on colors. In this paper more focus area is the way of combination of clustering technique in order to get faster output of images. In this paper the clustering techniques are discussed and analyzed. Also, we propose a method HDK that uses more than one clustering technique to improve the performance of CBIR. This method makes use of hierarchical and divides and conquers K-means clustering technique with equivalency and compatible relation concepts to improve the performance of the K-Means for using in high dimensional datasets. It also introduced the feature like color, texture and shape for accurate and effective retrieval system.

**Keywords : -** Image, CBIR, Query, Retrieval, Clustering.

## I. INTRODUCTION

Interest in the potential of digital images has increased enormously over the last few years, fuelled at least in part by the rapid growth of imaging on the World-Wide Web. Users in many professional fields are exploiting the opportunities offered by the ability to access and manipulate remotely-stored images in all kinds of new and exciting ways. However, they are also discovering that the process of locating a desired image in a large and varied collection can be a source of considerable frustration. The problems of image retrieval are becoming widely recognized, and the search for solutions an increasingly active area for research and development.

Problems with traditional methods of image indexing have led to the rise of interest in techniques for retrieving images on the basis of automatically-derived features such as color, texture and shape – a technology now generally referred to as Content-Based Image Retrieval (CBIR). After a decade of intensive research, CBIR technology is now beginning to move out of the laboratory and into the marketplace, in the form of commercial products like QBIC and Virage. However, the technology still lacks maturity, and is not yet being used on a significant scale. In the absence of hard evidence on the effectiveness of CBIR techniques in practice, opinion is still sharply divided about their usefulness in handling real-life queries in large and diverse image collections. Nor is it yet obvious how and where CBIR techniques can most profitably be used.

The process of digitization does not in itself make image collections easier to manage. Some form of cataloguing and indexing is still necessary – the only difference being that much of the required information can now *potentially* be derived automatically from the images themselves.

The need for efficient storage and retrieval of images – recognized by managers of large image collections such as picture libraries and design archives for many years – was reinforced by a workshop sponsored by the USA's National Science Foundation in 1992. After examining the issues involved in managing visual information in some depth, the participants concluded that images were indeed likely to play an increasingly important role in electronically-mediated communication. However, significant research advances, involving collaboration between a numbers of disciplines, would be needed before image providers could take full advantage of the opportunities offered. They identified a number of critical areas where research was needed, including data representation, feature extractions and indexing, image query matching and user interfacing.

Access to a desired image from a repository might thus involve a search for images depicting specific types of object or scene, evoking a particular mood, or simply containing a specific texture or pattern. Potentially, images have many types of attribute which could be used for retrieval. This leads naturally on to a classification of query types into three levels of increasing complexity.

***Level 1*** - comprises retrieval by *primitive* features such as color, texture, shape or the spatial location of image elements.

***Level 2*** - comprises retrieval by *derived* features, involving some degree of logical inference about the identity of the objects depicted in the image.

***Level 3*** - comprises retrieval by *abstract* attributes, involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted.

In recent years there had been significant growth in research into methods for organizing and retrieving similar images in the database. Some of the techniques include text labeling and Content Based Image Retrieval (CBIR). In text labeling each image is indexed by their keywords. The keywords are chosen in a way that best represents the image that is stored, so reducing the search to merely a keyword search which is done using some string matching techniques. One particular disadvantage of this approach is its supervisory nature. In CBIR images are retrieved automatically based on the color, texture and shape features of the image

## II. WHAT IS CBIR?

As processors become increasingly powerful, and memories become increasingly cheaper, the deployment of large image databases for a variety of applications have now become realizable. Databases of art works, satellite and medical imagery have been attracting more and more users in various professional fields — for example, geography, medicine, architecture, advertising, design, fashion,

and publishing. Effectively and efficiently accessing desired images from large and varied image databases is now a necessity CBIR differs from classical information retrieval in that image databases are essentially unstructured, since digitized images consist purely of arrays of pixel intensities, with no inherent meaning. One of the key issues with any kind of image processing is the need to extract useful information from the raw data before any kind of reasoning about the image's contents is possible. Image databases thus differ fundamentally from text databases.

CBIR or Content Based Image Retrieval is the retrieval of images based on visual features such as color, texture and shape. Reasons for its development are that in many large image databases, traditional methods of image indexing have proven to be insufficient, laborious, and extremely time consuming. These old methods of image indexing, ranging from storing an image in the database and associating it with a keyword or number, to associating it with a categorized description, have become obsolete. This is not *CBIR*. In CBIR, each image that is stored in the database has its features extracted and compared to the features of the query image. It involves two steps:

- Feature Extraction: The first step in the process is extracting image features to a distinguishable extent.
- Matching: The second step involves matching these features to yield a result that is visually similar.

Image retrieval in a larger database is quite complex process. For this efficient features must be extracted from the images for its training.

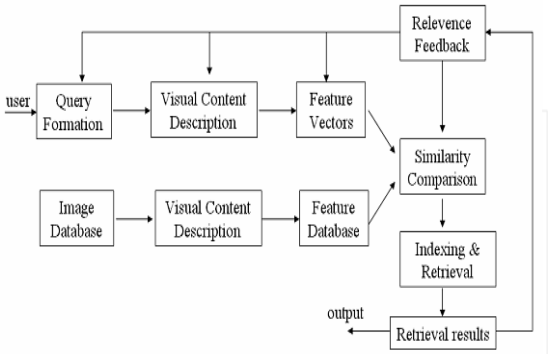


**Figure 1-1**. Diagram for content-based image retrieval system

## 2.1. CBIR Systems

Several CBIR systems currently exist, and are being constantly developed. Examples are:

- **QBIC** or **Query By Image Content** was developed by IBM, Almaden Research Centre, to allow users to graphically pose and refine queries based on multiple visual properties such as color, texture and shape. It supports queries based on input images, user-constructed sketches, and selected color and texture patterns.
- **VIR Image Engine** by Virage Inc., like QBIC, enables image retrieval based on primitive attributes such as color, texture and structure. It examines the pixels in the image and performs an analysis process, deriving image characterization features.
- **VisualSEEK** and **WebSEEK** were developed by the Department of Electrical Engineering, Columbia University. Both these systems support color and spatial location matching as well as texture matching.
- **NeTra** was developed by the Department of Electrical and Computer Engineering, University of California. It supports color, shape, spatial layout and texture matching, as well as image segmentation.
- **MARS** or **Multimedia Analysis and Retrieval System** was developed by the Beckman Institute for Advanced Science and Technology,

University of Illinois. It supports color, spatial layout, texture and shape matching.

**Viper** or **Visual Information Processing for Enhanced Retrieval** was developed at the Computer Vision Group, University of Geneva. It supports color and texture matching.

## III. LITERATURE REVIEW

Several reviews of the literature on image retrieval have been published, from a variety of different viewpoints.

- Enser [1995] reviews methods for providing subject access to pictorial data, developing a four-category framework to classify different approaches. He discusses the strengths and limitations both of conventional methods based on linguistic cues for both indexing and search, and experimental systems using visual cues for one or both of these. His conclusions are that, while there are serious limitations in current text-based techniques for subject access to image data, significant research advances will be needed before visually-based methods are adequate for this task. He also notes, as does Cawkell [1993] in an earlier study, that more dialogue between researchers into image analysis and information retrieval is needed.
- Aigrain et al [1996] discuss the main principles of automatic image similarity matching for database retrieval, emphasizing the difficulty of expressing this in terms of automatically generated features. They review a selection of current techniques for both still image retrieval and video data management, including video parsing, shot detection, keyframe extraction and video skimming. They conclude that the field is expanding rapidly, but that many major research challenges remain, including the difficulty of expressing semantic information in terms of primitive image features, and the need for significantly improved user interfaces. CBIR techniques are likely to be of most use in restricted subject domains, and where synergies with other types of data (particularly text and speech) can be exploited.
- Eakins [1996] proposes a framework for image retrieval (outlined in section **Error! Reference source not found.** above), classifying image queries into a series of levels, and discussing the extent to which advances in technology are likely to meet users' needs at each level. His conclusion is that automatic CBIR techniques can already address many of users' requirements at level 1, and will be capable of making a significant contribution at level 2 if current research ideas can be successfully exploited. They are however most unlikely to make any impact at level 3 in the foreseeable future.
- Idris and Panchanathan [1997a] provide an in-depth review of CBIR technology, explaining the principles behind techniques for color, texture, shape and spatial indexing and retrieval in some detail. They also discuss the issues involved in video segmentation, motion detection and retrieval techniques for compressed images. They identify a number of key unanswered research questions, including the development of more robust and compact image content features, more accurate modelling of human perceptions of image similarity, the identification of more efficient physical storage and indexing techniques, and the development of methods of recognizing objects within images.
- De Marsicoi et al [1997] also review current CBIR technology, providing a useful feature-by-feature comparison of 20 experimental and commercial systems.
- In addition to these reviews of the literature, a survey of "non-text information retrieval" was carried out in 1995 on behalf of the European Commission by staff from GMD (Gesellschaft für

Mathematik und Datenverarbeitung), Darmstadt and Université Joseph Fourier de Grenoble [Berrut et al, 1995]. This reviewed current indexing practice in a number of European image, video and sound archives, surveyed the current research literature, and assessed the likely future impact of recent research and development on electronic publishing. The survey found that all current operational image archives used text-based indexing methods, which were perceived to have a number of shortcomings. In particular, indexing vocabularies were not felt to be adequate for non-text material. Despite this, users seemed generally satisfied with existing systems. The report concluded that standard information retrieval techniques were appropriate for managing collections of non-text data, though the adoption of intelligent text retrieval techniques such as the inference-based methods developed in the INQUERY project [Turtle and Croft, 1991] could be beneficial.

## IV. PROBLEM STATEMENT

Image databases and collections can be enormous in size, containing hundreds, thousands or even millions of images. The conventional method of image retrieval is searching for a keyword that would match the descriptive keyword assigned to the image by a human categorizer. Currently under development, even though several systems exist, is the retrieval of images based on their content, called Content Based Image Retrieval, CBIR. While computationally expensive, the results are far more accurate than conventional image indexing. Hence, there exist tradeoffs between accuracy and computational cost. This tradeoffs decreases as more efficient algorithms are utilized and increased computational power becomes inexpensive.

### 4.1. Problem Statement

The problem involves entering an image as a query into a software application that is designed to employ CBIR techniques in extracting visual properties, and matching them. This is done to retrieve images in the database that are visually similar to the query image.

### 4.2. Proposed Solution

The solution initially proposed was to extract the primitive features of a query image and compare them to those of database images. The image features under consideration were colour, texture and shape. Thus, using matching and comparison algorithms, the colour, texture and shape features of one image are compared and matched to the corresponding features of another image. This comparison is performed using colour, texture and shape distance metrics. In the end, these metrics are performed one after another, so as to retrieve database images that are similar to the query. The similarity between features was to be calculated using algorithms used by well known CBIR systems. For each specific feature there was a specific algorithm for extraction and another for matching.

## V. STUDY AND WORKFLOW OF THE PROPOSED SYSTEM

Content-based image retrieval uses the visual contents of an image such as *color, shape,* and *texture* to represent and index the image. To retrieve images, users provide the retrieval system with example images or sketched figures. The system then changes these examples into its internal representation of feature vectors. The similarities /distances between the feature vectors of the query example or sketch and those of the images in the database are then calculated and retrieval is performed.

The indexing scheme provides an efficient way to search for the image database. Recent retrieval

systems have incorporated users' relevance feedback to modify the retrieval process in order to generate perceptually and semantically more meaningful retrieval results.

The study of the proposed system can be categorized into three modules:-

- **Module 1:-** Content based Image Retrieval using HOG of Wavelet subbands.
- **Module 2:-** A New Generalized Gaussian Density and HOG based content image retrieval system.
- **Module 3:-** A New content based image retrieval using curvelet Transform.

CBIR differs from classical information retrieval. In that approach, image databases are primarily unstructured, since digitized pictures consist strictly of arrays of pixel intensities, with no inherent. One of the key problems associated with any reasonably image process is that they got to extract helpful information from the large volume of available information before concerning whether the image's contents is feasible.

## VI. CONCLUSION

The dramatic rise in the sizes of images databases has stirred the development of effective and efficient retrieval systems. The development of these systems started with retrieving images using textual connotations but later introduced image retrieval based on content. This came to be known as CBIR or Content Based Image Retrieval. Systems using CBIR retrieve images based on visual features such as colour, texture and shape, as opposed to depending on image descriptions or textual indexing. In this project, we have researched various modes of representing and retrieving the image properties of colour, texture and shape. Due to lack of time, we were only able to fully construct an application that retrieved image matches based on colour and texture only.

The application performs a simple colour-based search in an image database for an input query image, using colour histograms. It then compares the colour histograms of different images using the Discrete Wavelet Transform (DWT). Further enhancing the search, the application performs a texture-based search in the colour results, using wavelet decomposition and energy level calculation. It then compares the texture features obtained using the Generalized Gaussian density (GGD). A more detailed step would further enhance these texture results, using a shape-based search.

## VII. REFERENCES

[1]. H. Tamura, and N. Yokoya, "Image Database Systems: A Survey," Pattern Recognition, vol. 17, no 1, pp.29–49, Sep. 1984.

[2]. S. Gerard, C. Buckely, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management, vol. 24, no.5, pp. 513-523, Jan. 1988.

[3]. Y. Chen, J. Wang, "Image Categorization by Learning and Reasoning with Regions," Journal of Machine Learning Research, vol. 5, pp. 913–939, May 2004.

[4]. F. Long, H. Zhang, H. Dagan, and D. Feng, "Fundamentals of content based image retrieval," in D. Feng, W. Siu, H. Zhang (Eds.): "Multimedia Information Retrieval and Management. Technological Fundamentals and Applications," Multimedia Signal Processing Book, Chapter 1, Springer-Verlag, Berlin Heidelberg New York, 2003, pp.1-26.

[5]. V. Gudivada and V. Raghavan, "Content-based image retrieval systems," IEEE Computer, vol. 28, no 9, pp18-22, Sep. 1995.

[6]. M. Kherfi, D. Ziou, and A. Bernardi, " Image Retrieval From the World Wide Web: Issues,

Techniques, and Systems," ACM Computing Surveys, vol. 36, no. 1, pp. 35–67, March 2004.

[7]. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, and P. Yanker, "Query by image and video content:The QBIC system," IEEE Computer, vol. 28, no 9, pp.23-32, Sep. 1995.

[8]. A. Pentland, R. Picard, and S. Sclaroff , "Photobook: Content based manipulation of image databases," International Journal of Computer Vision, vol.18, no 3, pp.233–254, June 1997.

[9]. J. Smith and S. Chang, "Visualseek: A Fully Automated Content-Based Image Query System," Proceedings of the 4th ACM international conference on Multimedia table of contents, Boston, Massachusetts, United States, Nov. 1996, pp. 87-98.

[10]. A. Gupta, and R. Jain, "Visual information retrieval," Comm. Assoc. Comp. Mach., vol. 40, no. 5, pp. 70–79, May. 1997.

[11]. J. Li, J. Wang, and G. Wiederhold, "Integrated Region Matching for Image Retrieval," In Proceedings of the 2000 ACM Multimedia Conference, Los Angeles, October 2000, pp. 147-156.

[12]. Y. Deng, B. Manjunath, "Unsupervised Segmentation of Color -Texture Regions in Images and Video," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 8, pp. 800-810, Aug. 2001.

[13]. A. Smeulders, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349- 1380, May. 2000.

[14]. J. Caicedo, F. Gonzalez, E. Romero, E. triana, "Design of a Medical Image Database with Content-Based Retrieval Capabilities," In Proceedings of the 2nd Pacific Rim conference on Advances in image and video technology, Santiago, Chile, December 17-19, 2007.

[15]. B. Manjunath and W. Ma, "Texture features for Browsing and retrieval of image data," IEEE transactions on pattern analysis and machine intelligence, vol. 18. No. 8, pp. 837-842, August 1996

[16]. R. Zhang, and Z. Zhang, "A Clustering Based Approach to Efficient Image Retrieval," Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02), Washington, DC, Nov. 2002, pp. 339-346.

**Cite this article as :**

Dr. S. Dhinakaran, "Study of Content Based Image Retrieval Using Data Mining Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.397-403, May-June-2020. Available at
doi : https://doi.org/10.32628/CSEIT206385
Journal URL : http://ijsrcseit.com/CSEIT206385