Color –Based Image Retrieval in Image Database System

Gunja Varshney, Uma Soni

Abstract-Image Databases (IDBs) are a special kind of Spatial Databases where a large number of images are stored and queried. IDBs find a plethora of applications in modern life, e.g. in Medical, Multimedia, Educational Applications, etc. Data in an IDB may be stored in raster or vector format. Each of these data formats has certain properties and, in several cases, the choice between them is a challenge. Raster data lead to fast computing of several operations and they are well suited to remote sensing. On the other hand, they have a fixed resolution, leading to limited detail. In this article, we focus on raster data. We present the design and architecture of an Image Database System where several query types are supported. These include: queries about the additional properties (descriptive information) that have been recorded for each image (e.g. which images have been used as *covers of children's books), queries about the color characteristics* (color features) of the images (e.g. find the images that depict vivid blue), queries by example, or sketch (e.g. a sample image is chosen, or drawn by the user and images color-similar to this sample are sought). Color retrieval is achieved by utilizing color histograms. The development of our system is based on non-specialized tools: a relational database, Visual Basic and the *computer's file system. The user inter*face of the system aims at increased ease of use. It permits the management of the collection of images and the effective querying of the images by all the above query types and their combinations.

Keywords: Image Databases, Color Information, Query Processing, Color-Based Retrieval, Spatial Databases, Image Retrieval

I. INTRODUCTION

Image Databases (IDBs) are a special kind of Spatial Databases where a large number of images are stored and queried. IDBs find a plethora of applications in modern life, e.g. in Medical, Multimedia, Educational Applications, etc.

What makes an IDB a useful tool is its ability to answer queries, like the following [9]:

Manuscript received July 09, 2011.
Gunja Varshney is with Computer Science & Engineering, Sobhasaria Engineering College, Sikar , Rajsthan , India (e-mail: gunja.cs@gamil.com).
Uma Soni is with Computer Science & Engineering, Sobhasaria Engineering College, Sikar, Rajsthan, India (e-mail:11umasoni@gamil.com).

- Queries about the content of additional properties (descriptive information) that have been embedded for each image (e.g., which images have been used in the book cover of children's books?).
- Queries about the characteristics/features of the images like color, texture, shape etc. (e.g., find the images that depict vivid blue sky.).
- Queries for retrieving images with specified content (e.g., find the images that contain the sub-image of a specified chair.).
- Queries by example or sketch (e.g., A sample image is chosen, or drawn by the user and images similar to this sample are sought.).
- Structural queries (e.g., find the images that contain a number of specific objects in a specified arrangement.).
- Image Joins (e.g., find the cultivation areas that reside in polluted atmosphere areas.).
- Queries that combine regional data and other sorts of spatial data (e.g., find the cities represented by point data that reside within 5km from cotton cultivations.).
- Temporal Queries on sequences of evolving images (e.g., find if there has been an increase in the regions of wheat cultivations in this prefecture during the last two years.).

The importance of image querying led major manufacturers of Database Management Systems to embed related extensions to the core engine of their products, e.g. DB2 (QBIC) and Oracle (Virage).

Data in an IDB may be stored in raster or vector format. Each of these data formats has certain properties and, in several cases, the choice between them is a challenge. Raster data lead to fast computing of several operations and they are well suited to remote sensing. On the other hand, they have a fixed resolution, leading to limited detail.

In this article, we focus on IDBs storing raster data. We present the design and architecture of an IDB System where several query types are supported. These include: queries about the additional properties (usually called text-based image retrieval) that have been recorded for each image, queries about the color characteristics / features of the images, queries by example, or sketch (images color-similar to the sample image are sought). Color retrieval is achieved by utilizing color histograms. The development of our system is based on non-specialized tools: a relational database,

Visual Basic and the computer's file system. The user interface of the system aims at increased ease of use. It permits the management of the collection of images and the effective querying of the images by all the above query types and their combinations.

The rest of this article is organized as follows. In Section 2, we briefly review content-based image retrieval and refer to several existing IDB systems. In Section 3, we present an introduction to color-based image retrieval. In Section 4, we present the architecture and retrieval techniques of our system. In Section 5, we present the user interface and use of our system. In Section 6, we present an evaluation of the use of our system. Finally, in the last section we conclude the contribution of our work and present possible future research directions.

II. CONTENT-BASED IMAGE RETRIEVAL

More effective techniques than simple browsing are necessary for searching collections of large numbers of images. An initial approach for organizing such image collections is to use words that refer to properties of the image, such as the creator, the place, the time, or the subject of the image. The technique that is based on words for image retrieval is called text-based image retrieval or metadata-based image retrieval and constitutes a traditional technique that has been used during previous times for analog image collections.

The technique for image retrieval from a digital collection by using feature-element values that are extracted automatically from the optical contents of the images is called content-based image retrieval. Feature extraction and analysis is performed from the images so that resulting values are comparable by the use of a computing machine for examining the similarity between images. Useful features for content-based image retrieval are considered those that mimic the features seen by humans, those that are perceived by the human vision. The use of such optical features, that reflect a view of image similarity as this is perceived by a man, even if he has difficulty in describing these features, increases the probability that the system recalls images that are similar, or alike, according to the human perception.

The features that are used for content-based image retrieval are characterized as global (local) when they refer to the whole (a part of the) image. The basic characteristics that are used for content-based image retrieval are: the color (the distribution, or analogy of different colors at parts, or the whole image), the shape (the shape of the boundaries, or the interiors of objects depicted in the image), the texture (the presence of visual patterns that have properties of homogeneity and do not result from the presence of single color, or intensity), the location (the relative to other objects, or absolute position where each object resides in the image).

Several systems have been developed for content-based image retrieval. Some of the most well known are: ALISA, Blobworld, CANDID, CHROMA, COMPASS, Excalibur Visual RetrievalWare, FIDS, FIR, ImageRETRO, ImageRover, iPure, KIWI, MARS, Metaseek, Photobook, PicSOM, QBIC, SMURF, SIMPLIcity, FIRE. More details regarding these systems appear in [3, 5]. More details regarding content-based image retrieval techniques and achievements appear in [1, 2, 6, 7].

III. COLOR-BASED IMAGE RETRIEVAL

Figures digital image may be considered as a two dimensional array where the array cells correspond to the image pixels and the values stored in the cells to the values of color-intensity, in case of a grayscale (single-color) image. A color image consists of three single-color images that correspond to the colors Red, Green and Blue (from which any color may be composed, when appropriate intensity values are combined). By making a function from the discrete values of intensity to the number of pixels with the respective value, we construct a Histogram for each of the component colors [4]. By grouping together several neighboring values of intensity, we decrease the number of histogram bins and thus limit the number of required calculations.

The method that we use in our system for color-based image retrieval is based on checking of the similarity between histograms. This can be accomplished by a method called Histogram Intersection [4]. First of all, independence to the image size is required, so that histogram intersection of images with different sizes can be done. If H(i) is the histogram of an image, where i represents a histogram bin, then the normalized histogram is defined as:

$$I(i) = \frac{H(i)}{\sum_i H(i)}$$

By considering that I_R, I_G and I_B are the normalized histograms of an image of our database and Q_R, Q_G και Q_B the normalized histograms of the image to be searched (for the colors red, green and blue, respectively), the similarity between two images is given by the following formulae:

$$S_R = \frac{\sum_R \min\left(I_R(r), Q_R(r)\right)}{\sum_R Q_R(r)}$$

$$S_G = \frac{\sum_G \min\left(I_G(g), Q_G(g)\right)}{\sum_G Q_G(g)}$$

$$S_B = \frac{\sum_B \min\left(I_B(b), Q_B(b)\right)}{\sum_B Q_B(b)}$$

$$S_C(I,Q) = \frac{S_R + S_G + S_B}{3}$$

If the histograms are identical (similar) then S_c=1 (S_c~1). The basic advantage of this method is that the color histograms are independent to rotation and translation, since the color

similarity is calculated, without information about its spatial distribution.

IV. SYSTEM ARCHITECTURE & QUERY PROCESSING

The architecture of our content-based retrieval system appears in the following figure.

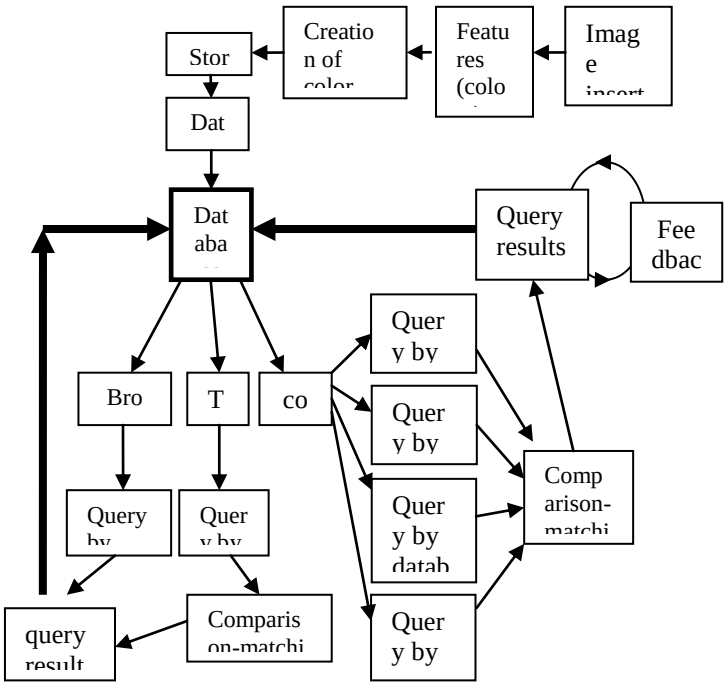

Figure 1: the architecture of our content-based retrieval system

Note that for text / properties-based search the user fills one or more of the following fields, referring to an image: Filename, keywords, resolution, color depth, file size, file path, receipt date, store date, compression type, subject, creator, comment. In case more than one field is filled, the search is performed based on a logical and between them.

For color-based search, the user is able to query the database by using as a query prototype the current database image, a new image, an image sketch created by himself, or by defining the distribution (percentages) of the three constituent colors (red, green, blue).

The results of a color-based search appear in decreasing percentage of similarity that appears next to the name of each image. At this stage, the user may pick an image from the query result and use it as a prototype for performing a new search (relevance feedback.

V.USER INTERFACE

In this section some characteristic (among the numerous) screenshots of our system user interface are presented. Note that the user interface is currently only in Greek.

In Figure 2, the initial administration scheme of our system (a starting point for all the basic system functions) is depicted. From this screen, the user may chose a function, or browse the images in the database one-by-one (the first image of the database is initially shown). In Figure 3, the screen that permits text / properties-based search is depicted.

Figure 2: the initial administration screen

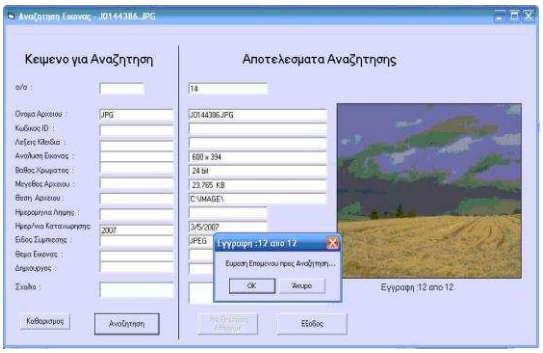

Figure 2: Text Properties based Search

In Figure 4, the screen that permits the user to choose a color-based search type is depicted (by using the current database image, a new image, an image sketch created by himself, or by defining the distribution of the constituent colors). In Figure 5, the result of a color-based search is depicted. The user may choose any of the retrieved images

and perform a new search based on it (relevance feedback). In Figure 6, the results of search by browsing (in group format) are depicted

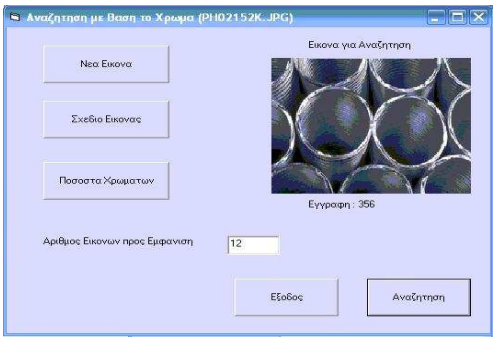

Figure 4: color-based search type choice

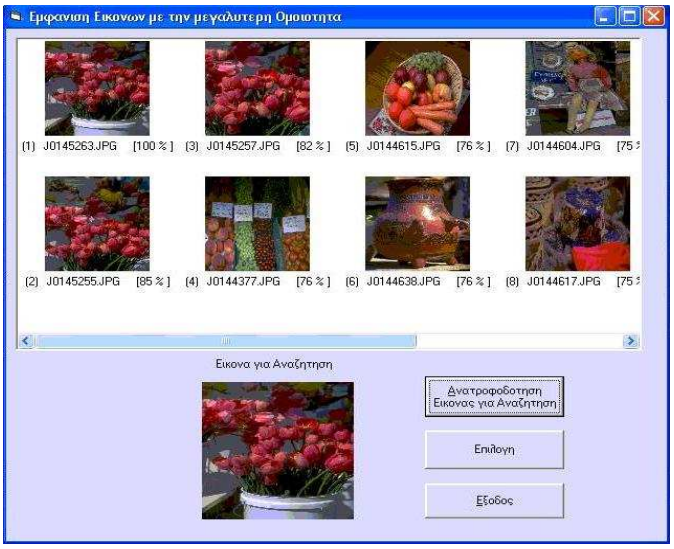

Figure 5: results of a color-based search

Figure 6: results of a browsing search

VI. EVALUATION OF SYSTEM USE

The performance tests of our system took place on a desktop computer, as well as on a (less powerful) laptop computer. The results that will be presented in the following come from the laptop computer (Intel M Cetrino 1,73 GHz CPU, 512 MB RAM, 60 GB Hard Disk). We created a database with images covering a wide range of scenes (landscapes, objects, faces, etc), resolutions and color distributions. We performed several tests for each search and we calculated average search times.

It should be noted that the text / properties-based search is very small even for a large number of images (practically zero). For color-based search, the search time is always the same for the same database (same number of images) and independent to the size, visual complexity and resolution of the database images, as well as, of the query image. This is due to the fact that for searching the whole database must be scanned and the similarity factors for each image need to be calculated. Note that, for the sake of performance, in the database histograms for downscaled versions of the original images are stored (128x128 pixels). This was found not to have a significant result on search accuracy.

In Table 1, we depict the search times for several cardinalities of the image set. Note that after the search, the similarity factors are sorted in decreasing order, followed by the respective position pointers of the images in the database (the image records themselves are not sorted). The resulting search times show that even in the case of the small computing system used, the performance is acceptable for a wide range of practical applications. The quality of the retrieval result, according to subjective human perception of the system testers, was quite good and images that interested the testers were successfully located.

# images	530	900	1800	3500	4500
Time (sec)	1,5	2,4	3	3,9	5,3

Table 1: color-based search performance

VII. CONCLUSIONS

In this article, we presented a system for content-based image retrieval that supports search by text / properties and search by color contents. This system is capable of / supports

- automatic storage of certain image properties, like image size and resolution,
- batch insertion of several images (and directories containing image files),
- no limitations regarding the (local, or network) paths of image files and the type of image compression,

- a database of small footprint, since only the histograms of images are stored in the database, while the image files remain on the file system,
- multiple search-by-color functions,
- user friendliness though a self-explained graphical user interface with graphical representation of images at each step of the system use,
- on-line help,
- tools for the administration of the database,
- the use of multiple image collections,
- the capability to manipulate the database using SQL through the graphical user interface.

Future research plans include the addition of:

- search by color, based on local color information (of specific image parts),
- search by shape, or texture,
- capability of combined search by text / properties, color, shape and/or texture,
- improvement of the search times by making use of specialized indices [8, 9],
- search on the world-wide-web,

Web-based access to the system

REFERENCES

1 Athanasakos K. C., Doulamis A. D. and Karanikolas N. N. "A Signature Tree Content-based Image Retrieval System", Proc. 3IA'2007 - 10th International Conference on Computer Graphics and Artificial Intelligence, May 30-31, 2007, Athens, Greece, pp. 181-191

2 Datta R., Li J., Wang J.Z., (2005) "Content-Based Image Retrieval - Approaches and Trends of the New Age," Proceedings of the 7th International Workshop on Multimedia Information Retrieval, in conjunction with ACM International Conference on Multimedia, Singapore, pp. 253 - 262

3 Johansson, B. (2000) "A survey on: Contents Based Search in Image Databases", technical report LiTH-ISY-R-2215, Linköping University

4 Long F., Zhang H.J., and Feng D.D. (2003) "Fundamentals of Content-Based Image Retrieval", Multimedia Information Retrieval and Management- Technological Fundamentals and Applications, D. Feng, W.C. Siu, and H.J. Zhang (Eds.), Springer, 2003

5 Moutousidis E., A System for Content-Based Image Retrieval, Master's thesis submitted for approval, Helenic Open University (2007).

6 Rui Y., Huang T.S. and Chang S.-F. (1999), "Image Retrieval: Current Techniques, Promising Directions and Open Issues", Journal of Visual Communication and Image Representation, 10(1): p. 39-62

7 Schettini, R., Ciocca, G., & Zuffi, S. (2001). A survey of methods for colour image indexing and retrieval in image databases. In L. W. MacDonald & M. R. Luo (Eds.), Color imaging science: exploiting digital media. Chichester, England: Wiley, J. & Sons Ltd.

8 Valova I., Rachev B. & Vassilakopoulos M., Optimization of the Algorithm for Image Retrieval by Color Features, Proc. CompSysTech' 2006, II.17 1-5

9 Vassilakopoulos M., Corral A., Rachev B., Valova I. & Stoeva M., IMAGE DATABASE INDEXING TECHNIQUES, to appear in the Encyclopedia of Geoinformatics