Image Databases – An Approach for Image segmentation & Color reduction Analysis & Synthesis

Irena Valova, Boris Rachev

Abstract: *In this paper, we describe a system for image databases organization and retrieval based on the pictorial content representation. We suggest extending the use of image histograms to characterize the global and local color properties of an image and examine the effect of the degree of image segmentation and a number of known color reduction algorithms on the time taken for image database creation and retrieval and on the size of the metadata, extracted from the images. We drew the following conclusions from the analysis of the results of our experiments: the color reduction algorithms have a relatively small effect while the size of the database and the times taken for creation and retrieval increase significantly if the image segmentation is above 16X16. We have also concluded that this type of structure should be based on a hierarchical method of segmentation.*

Key words: *image databases, organization, retrieval, architecture of image database management system.*

INTRODUCTION

A key issue in the design of image database system is the development of effective techniques for the automatic, unsupervised representation of pictorial content. In the last few years, several approaches based on the automatic extraction of image representations have been proposed, based on color and texture, shape, spatial relationships, and combinations of these features [1, 2, 3, 4]. A widespread representation of image color content uses a color clustering technique based on a single color histogram giving the distribution of pixels of each color in the color space of the image. Beside being effective for characterizing the *global* color properties of an image, the color histogram can also be used to define a measure of similarity between two images: the "histogram intersection" operator introduced in [5] provides a simple way to match two images through their color histograms.

In this work we will focus our attention on the color properties of images, introducing and analyzing a new effective method for the organization of image databases.

DEFINITION OF THE PROBLEM

The image retrieval process can be divided in two steps:

- **Indexing -** for each image in a database a set or a vector of features summarizing its content properties is computed and stored in the metadata database;
- **Retrieval** – given a query image its features are extracted and compared to the others in the database. Database images are then ordered following a similarity criterion

Color is by far the most common visual feature used in CBIR, primarily because of the simplicity of extracting color information from. Color histograms describe the distribution of pixels of each color in the color space of the image.

As a result of our research we found that we need a ***new and more effective method for storage and retrieval in very large databases of realistic images, based on the global and local color features of the images. The basic parts in this method are color reduction algorithms and image segmentation rate and it is necessary to analyze their effect on the metadata database size and on the times for it creation and retrieval.***

PREVIOUS WORK

Color histograms are widely used to compare images because they are trivial to compute and robust with respect to rotation and small camera viewpoint variations. Example of the use of histograms can be found in most of the well known content-based

image retrieval systems like QBIC developed by IBM [8], Mars [9], Virage [10], ImageRover [11], and Netra [12].

Unfortunately classical histograms lose spatial information about pixels arrangements, so very different images can have similar color distributions. To solve this problem many indexing techniques include a va riable amount of spatial information.

Stricker and Dimai [13] divide an image into five partially overlapped regions, the central of which is elliptical. For each region, they compute the first three moments (average color, variance and skewness) from each color histogram. They use a similarity function that considers the possibility of 90° image rotations.

Smith and Chang [14] partition an image in regions using a sequential labeling algorithm based on the selection of a single color or a group of colors. For each region they compute a binary color set using histogram back projection.

Pass and Zabih [15] described a split histogram called a color coherence vector. Each one of its buckets j contains pixels having a given color and two classes based on the pixels spatial coherence.

Huang et al. [16] described the use of color correlograms to integrate color and spatial information.

SOLUTION OF THE PROBLEM

In our research in image databases [17], we propose a method for extending the use of image histograms to characterize the *local* and *global* color properties of images and better preserve their intrinsic geometric information. This method is based on two types of color representations:

Color Descriptor – to represent global color features of the images;

Color Descriptor Matrix – to represent spatial color features of the images.

To generate these representations we use two processes: color reduction and image segmentation.

Because it is impossible to classify all the colors and to store the color for every pixel from the image, it is necessary to use color reduction to represent the color image content that effective and efficient computation of color indexes usually demands. This is generally achieved by color space quantization, using a predefined model for color representation. Formally if C is a color space, and $P = \{c_1, c_2, ..., c_i, ..., c_n \mid c_i \in C, n << \|C\|\}$, a subset of C called the model for color representation. A function Q that maps each color in C to an element in P is called the quantizer, and is defined as: $Q : C \rightarrow P$. Color reduction is the process of finding an acceptable set of colors that can be used to represent the original colors of a digital image. It is a reductive compression technique that transforms a set containing a large number of colors to a set that contains a relatively small number of colors. The ultimate goal of any color-reduction algorithm is to minimize the perceptual differences between the original and quantized images. There are many different algorithms for color quantitization and in our system we use two of them - Median Cut and Popularity algorithm, with number of colors n=16. A description and comparison of these algorithms is shown in [6].

The term image segmentation refers to the partition of an image into a set of regions that cover it. In the presented method we suggested to divide the entire image into blocks. For each of the blocks, it has its own set of histograms indicating the distribution of a color's property inside the block. If we use more blocks to cover the image we will have more precision information about the colors in the image, but this will increase the database size.

As shown in figure 1, the interactive image database management system is composed of two main subsystems: one devoted to internal data management operations (**retrieval subsystem**) and the other (**interaction subsystem**) to the dialogue with the user. A **representation** engine that is embedded in the retrieval subsystem automatically

produces a description of an input image or query image based on its color content. New image descriptions can be added to system at storage time (blue arrows), and become part of the database index. At retrieval time (red arrows), the query description is matched against the index descriptions based on a suitable metric, and a re-ordering of the database is obtained, where database images are ranked in ascending order of similarity with the query image. The query image (external query) produced by the graphic composition engine directly reflects the user's current retrieval task. The user is also provided by the interaction subsystem with a visual feedback of both query image and system output: these can be used to update (refine, modify, etc.) the task.

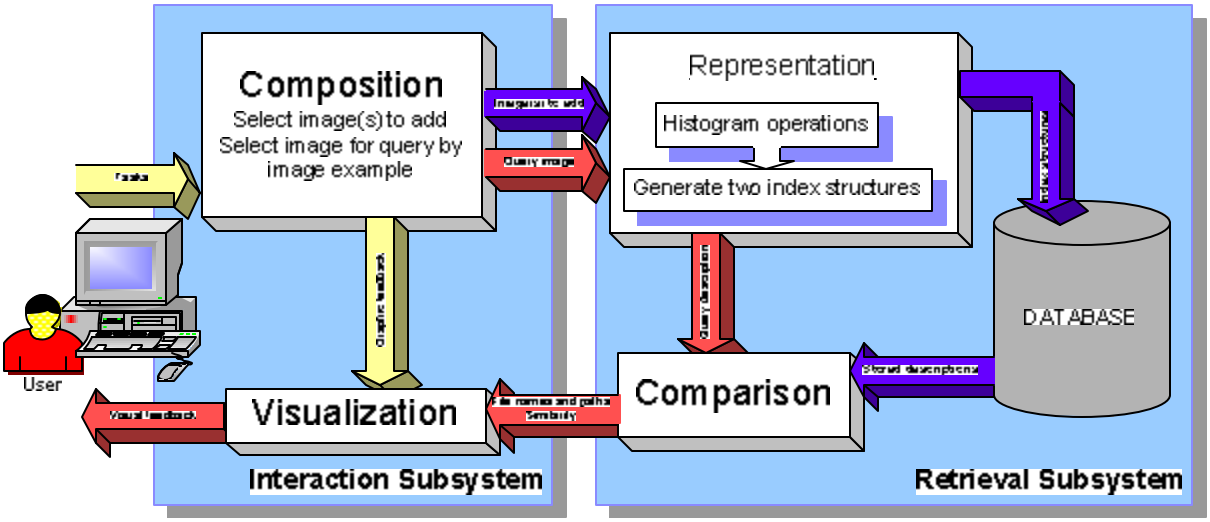


Figure 1. Architecture of the image database management system

Although quite straightforward to compute and to use, global color histograms are simply inadequate to represent the local characteristics of image content which are totally lost in the color clustering process. To retain image space information, color-based clustering must be extended to color-based segmentation.

The processes of the storage or query image representation are implemented in two stages.

The first stage of the representation is to generate the histogram of the selected colors for the image. The histogram is a list of "bins" showing the number of pixels being classified into the different color groups. To store the color image features we propose two types of color descriptors: dominant color descriptor and descriptor matrix. .

After color clustering, only a small number of colors remain and the percentages of these colors are calculated. Each representative color and its corresponding percentage form a pair of attributes that describe the color characteristics in an image. The **color descriptor** is defined to be

$F = \{\{c_i, p_i\}, i = 1,..., M\}$ where ? is the total number of color clusters in the image, c_i is a 3-D color vector, p_i is its percentage, and $\sum_i = p_i = 1$. Note that can vary from region to region.

This structure can store the color features of the images but it does not contain any spatial information. Due to these limitations which are significant for some type of queries we propose another structure – **Color Descriptor Matrix.**

In order to create this structure the whole image is divided into NxN (where N=4, 8, 16, 32) equal parts. This matrix stores the pairs of representative color and its corresponding percentage for the image blocks. The original images were NxN quantised and were represented as NxN blocks (or subimages).

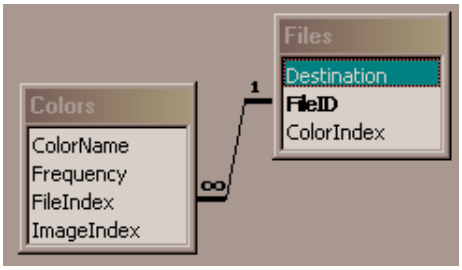


Figure 2. The storage of the metadata is in a relational database structure

Database - Image database consists of two data types – metadata, extracted from images (color distribution) and the images themselves. The storage of the metadata is in a relational structure, based on and supported by currently DBMS. Such design provides the potential benefits of using the existing features of Relational DBMS, such as the use of SQL (Structured Query Language) (Figure 2). In our experiments when we examine the database size we use this database of metadata.

The storage of the image data currently preserves the original "file-based" formats, and the original "directories" (or "folders"), which may be arranged by users themselves. Specifically, it means that it is based on "image files".

SOFTWARE IMPLEMENTATION, EXPERIMENTS AND RESULTS

We develop an image database management system with the architecture and functionality described above to study the effect of the segmentation rate and the color reduction algorithms on the metadata database size and on the creation and retrieval times (figure 3).

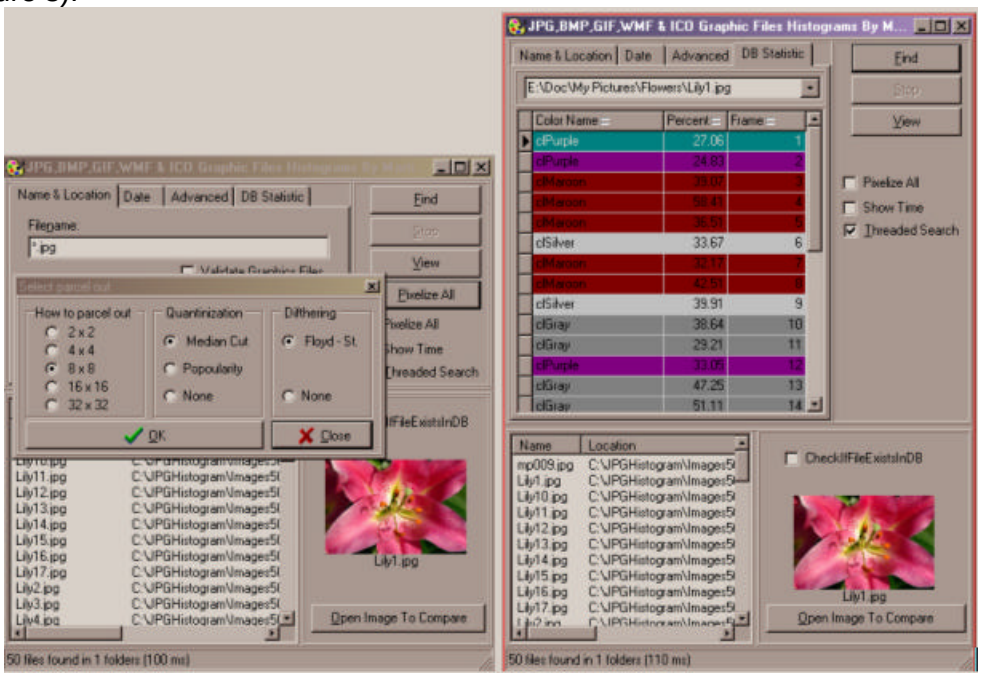


Figure 3. Image Database Management System

The system includes an easy to use optimized graphical user interface, and its functionality was divided into four main groups (see Figure 3).

Global color features queries -The user creates a query by choosing a sequence of colors (from the selected colors after reduction) in descending order according to its frequency in the requested images. Colors from the user's query are represented as its digital color descriptors and after that are transformed to SQL select query to the relational database.

Local color distribution queries - The image that is used as an example in the query is represented as local and global color features like all images in the database. As a result of these operations we have the same structures as we add the new image to the database. Next the Similarity field is calculated and the database is sorted by decreasing

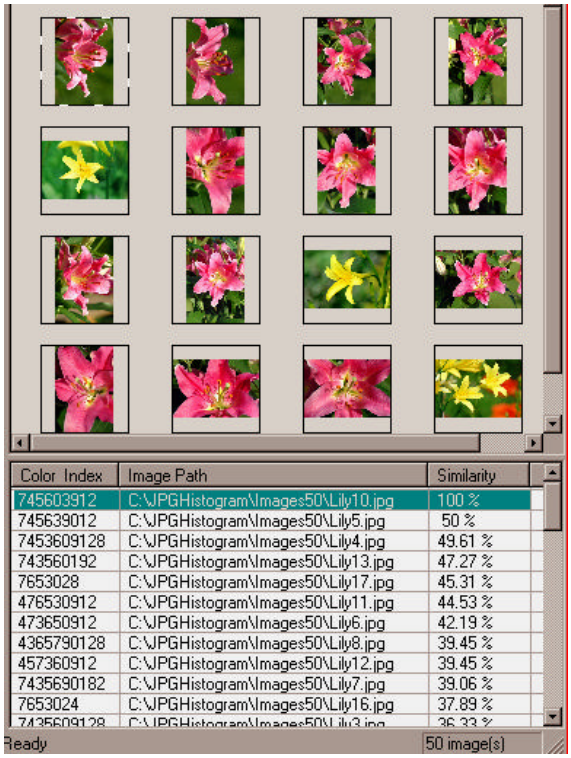


Figure 4. Local color distribution queries

of this field and visualized in the System Output Area and Feature Database Area windows (Figure 4).

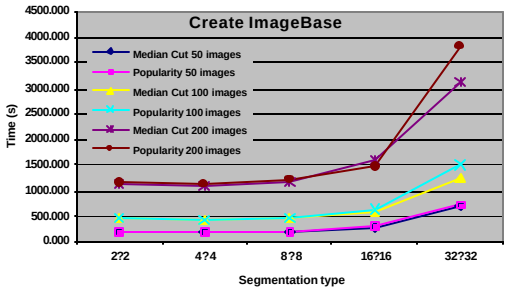
For our experiments we selected a database of 50, 100 and 200 images of natural scenes. Our tests were performed on a Celeron 300MHz based on PC under a Windows operating system. The program implementing the feature extraction, creation and retrieval processes was written in Microsoft Visual C++ 6.0.

We used RGB color space which is more classical than HSV, but it is possible to transform RGB to HSV, which better represents the human visual chromatic perception.

Images added to our databases were previously segmented to NxN blocks (N=4, 8, 16, 32), and a quantization was performed in order to reduce the colors number to 16 with respect to the selected algorithm (Median Cut or Popularity algorithms). In the last step, two structures for local and global color representation were extracted and stored.

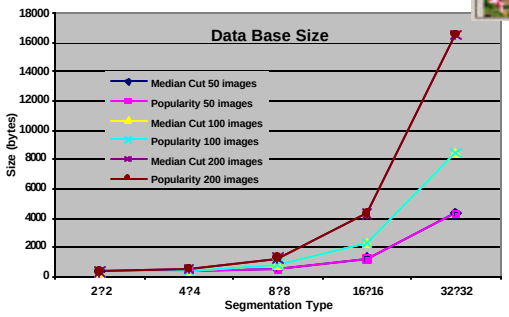
Table 1 Experimental results

		Color Reduction Algorithm	2?2	4?4	8?8	16?16	32?32
50 images	Creation time	Median Cut	205.020	191.460	201.240	300.690	686.560
		Popularity	200.930	190.070	205.670	305.930	739.850
	Retrieval time	Median Cut	21.862	22.763	30.935	55.890	129.457
		Popularity	50.964	23.965	31.365	60.128	158.801
	DB Size MB	Median Cut	268	316	511	1250	4310
		Popularity	268	316	511	1250	4310
100 images	Creation time	Median Cut	464.570	440.400	474.960	589.890	1243.360
		Popularity	463.010	441.170	472.320	627.130	1517.270
	Retrieval time	Median Cut	52.456	56.06	76.980	140.833	403.861
		Popularity	50.964	50.403	70.510	139.861	433.163
	DB Size MB	Median Cut	289	385	773	2260	8400
		Popularity	290	386	773	2260	8400
200 images	Creation time	Median Cut	1139.350	1116.160	1151.520	1587.650	3131.220
		Popularity	1175.570	1126.510	1211.000	1485.010	3824.150
	Retrieval time	Median Cut	148.943	154.612	224.503	403.600	967.902
		Popularity	144.147	156.545	232.394	401.878	932.872
	DB Size MB	Median Cut	333	527	1270	4280	16600
		Popularity	335	528	1270	4280	16600



a. Creation time

b. Retrieval time, example query with results



c. Database Size

Figure 5. Graphical interpretations of the experimental results from table 1

We perform the experiments with different number of images to create database and to retrieve all similar images to one selected image from the database. For the query by image example we use the same image in every separate experiment from our global experiment, but change only the number of images (50, 100 and 200 images) and segmentation rate (N=2,4,8,16,32) and color reduction algorithm (Median Cut or Popularity algorithms). We received the results shown in table 1 and visualized on figure 5 a, b, c.

CONCLUSIONS AND FUTURE WORK

In this paper, we have presented and discussed an interactive system for the organization of and retrieval from image databases based on color distributions and simple geometric properties of regions segmented from the images. A specific model for color representation was used in order to represent spatial color features of images.

After the analysis of the results from table 1 and figure 5 we can make the following conclusions: the creation and the retrieval time and the size of metadata database slightly depend on the chosen color reduction algorithm. As a result of the experiments with two algorithms – Median Cut algorithm and Popularity algorithm we decided not to make any other research for other color reduction algorithms.

While the color reduction algorithms have only a small effect, creation and the retrieval times are increasing in significant rate (two nearly three times, see table 1) if the image segmentation rate is above 256 (16X16).

As the segmentation rate of the images in the image database grows the precision of the queries grows too, but the segmentation rate more than 256 parts significantly increases the database size and the responding time (nearly four times).

It is possible to make the conclusion that this approach with local and global color features from the images is applicable in different areas but the segmentation rate should be selected in the dependence of the application.

These results allow us to create new methodologies for building more efficient Very Large Image Databases in respect of their structure and languages for image definition and retrieval. A structure like this should have the hierarchical segmentation method.

REFERENCES

1. Gupta A., Jain R. Visual Information Retrieval. Communications of the ACM, vol 40, n. 5, 1997.
2. Niblack W., et al. The QBIC Project: Querying Images by Content using Color, Texture and Shape. *Research Report 9203*, IBM Research Division, Almaden Research Center, 1993.
3. C. Colombo, I. Genovesi - DEA, Università di Brescia Image Querying and Retrieval by Multiple Color Distributions, *Workshop IMAGE AND VIDEO CONTENT-BASED RETRIEVAL*, February, 23rd 1998, http://www.itim.mi.cnr.it/Linee/Linea1/scaricare/workshop.htm
4. Chang S.K., Jungert E. Pictorial Data Management based upon the Theory of Symbolic Projections. *Journal of Visual Languages and Computing*, vol. 2, n. 2, 1991.
5. Swain M., Ballard D. Color Indexing. *, International Journal of Computer Vision*, vol. 7, n. 11, 1991.
6. http://algolist.manual.ru/graphics/quant/qoverview.php
7. Irena Valova, Boris Rachev, Image organization, querying and retrieval by color distribution features, International Conference CompSysTech'2002, Sofia, Bulgaria, 20-21 June 2002
8. http://www.qbic.almaden.ibm.com/
9. http://jadzia.ifp.uiuc.edu:8001/
10. http://www.virage.com/online/
11. http://vivaldi.ece.ucsb.edu/Netra
12. http://www.cs.bu.edu/groups/ivc/ImageRover/
13. M. Stricker and A. Dimai. "Color indexing with weak spatial constraints", *SPIE Proceedings*, volume 2670, pages 29-40, February 1996, IS&T/SPIE
14. J.R.Smith and S.Chang,. Tools and Techniques for Color Image Retrieval. In *Symposium on Electronic Imaging: Science and Technology – Storage & Retrieval for Image and Video Databases IV*, volume 2670, pages 426-237, San Jose CA, February 1996, IS&T/SPIE
15. G. Pass and R Zabih. "Histogram refinement for content-based image retrieval". *IEEE Workshop on Applications of Computer Vision*, pages 96-102, 1996.
16. J.Huang, S.R.Kumar, M.Mitra, Wei_Jing Zhu, R.Zabih. "Image indexing using color correlograms". In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 762-768, 1997.
17. Boris Rachev, Irena Valova, Silyan Arsov, An Approach for Image Organization and Retrieval in Realistic Image Databases. *7Th GIS Workshop, Potsdam Germany, June 2001*.

ABOUT THE AUTHORS

Irena Marinova Valova, University of Rousse, Phone:++359 82 888695, Irena@ecs.ru.acad.bg
Boris Rachev, TU – Varna, ???. 052 302431 (407), e-mail: RACHEV@ms.ieee.bg

Contents