Optimization of the Algorithm for Image Retrieval by Color Features

Irena Valova, Boris Rachev, Michael Vassilakopoulos

***Abstract** - In this paper, we describe our research in content based image retrieval systems, based on color features. The growing size of the database will result in long search time which may be unacceptable in many practical situations. We describe our approach to solving this problem and to improve the performance of the image database management system, based on color features, by using of sorted list structures.*

***Keywords**: Image databases, content based image retrieval, image features, color content*

INTRODUCTION

There is a rapid increase in the size of digital image collections together with the fast growth of the Internet in the recent years. Digital images have found their way into many application areas, including Geographical Information System, Office Automation, Medical Imaging, Computer Aided Design, Computer Aided Manufacturing, and Robotics.

For content-based image retrieval (CBIR), i.e. searching in image databases based on image content, several image retrieval systems have been developed. One of the first systems was the QBIC system [8]. Other popular research systems are BlobWorld [10], VIPER/GIFT [11], SIMBA [9], and SIMPLIcity [12]. All these systems compare images based on specific features in one way or another and therefore a large variety of features for image retrieval exists. Usually, CBIR systems do not use all known features as this would involve large amounts of data and increase the necessary computing time. Instead, a set of features appropriate to the given task is ususally selected, but it is difficult to judge beforehand which features are appropriate for which tasks. The fundamental idea of the CBIR approach is to generate automatically image descriptions directly from the image content by analyzing the content of the images. Such techniques are being developed by many research groups and commercial companies around the world.

DEFINITION OF THE PROBLEM

Color is the first and most straightforward visual feature for indexing and retrieval of images [5, 6, 7]. It is also the most commonly used feature in the field. In our previous work in the field of color content based image retrieval systems [1, 2, 3, 4] we proposed two types of image descriptors. First one is based on global color features of the images and represents the dominant colors in the images. It is used for hierarchical classification of the images in image database.

Another color descriptor, called ***ColorDescriptorMatrix***, we propose to describe local color features or color distribution in the images. The original images were NxN quantized and were represented as NxN blocks (or sub images). We examined the creation and retrieval time and database size depending on the size of N in [2] and made the conclusion that N=16 is the most adequate for our purposes. In order to create this index structure the whole image is divided into 256 equal parts. This matrix stores the coefficient of the dominant color from the selected color code book in the corresponding part of the image. The advantages and disadvantages of these two structures are presented in [1, 3]. Also in [3, 4] is described the main algorithm for image database organization and retrieval.

The growing size of the database will result in long search time which may be unacceptable in many practical situations. Even if the time required to compare two images is very short, the cumulative time needed to compare the query image with all database images is rather long and is probably longer than the time an average user wants to wait.

SOLUTION OF THE PROBLEM, EXPERIMENTS AND RESULTS

Our approach to solving this problem and to improve the performance of the image database management system, based on color features, is to use following list structures:

Add new images in image database

1. ***ColorDescriptorMatrix*** is calculated and determined. It consists of 256 elements per image.
2. 10 lists are formed per each of these 256 blocks. These lists consist of the image IDs. As an ID we use the consecutive number of the images in the database.

The number of lists is equal to 10 because after the implementation of the color reduction algorithms we have 10 possible dominant colors (from the color code book, presented in table 1). According to the dominant color in each block the image ID will be added to the corresponding lists of these colors and blocks.

Table 1

Color Descriptor	Color Mapped
0	Uncertain Colours: “very dark” or “very bright”
1 ⬜	White
2 ▩	Grey
3 ⬛	Black
4 🟥	Red, Pink
5 🟫	Brown, Dark Yellow, Olive
6 🟨	Yellow, Orange, Light Yellow
7 🟩	Green, Lime
8 🟦	Blue, Cyan, Aqua, Turquoise
9 🟪	Purple, Violet, Magenta

As a result of this approach ***every image from the image database will be added with its ID in 256 lists***. The total number of lists in the database will be: 256x10 = 2560 (*Figure* 1)

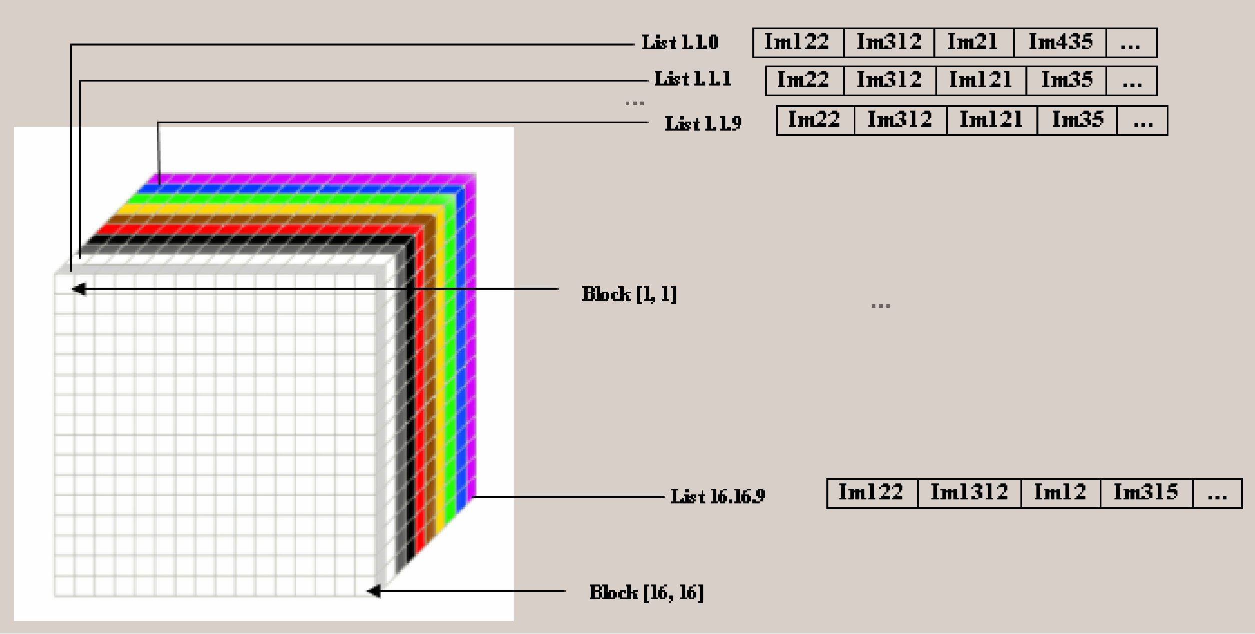

List 1.1.0
Im122 | Im312 | Im21 | Im435 | ...
List 1.1.1
Im22 | Im312 | Im121 | Im35 | ...
...
List 1.1.9
Im22 | Im312 | Im121 | Im35 | ...
Block [1, 1]
...
List 16.16.9
Im122 | Im1312 | Im12 | Im315 | ...
Block [16, 16]

Figure 1. An example list representation of the colors in quantized color images

Image retrieval

1. For both type of queries: ***Query by Image Example*** and ***Query by User Sketch*** first step is to calculate and define ***ColorDescriptorMatrix*** of the query, using the same algorithms as for adding new image in the database.
2. When we have the matrix of the query we analyze it consequently block by block and build temporal query list. This temporal query list contains the IDs from the corresponding lists in the image database of the dominant colors and the positions of the blocks in the query. So it contains 256 lists from the database.
3. The iterative IDs will be grouped after sorting of this temporal query list. The number of IDs in every group is determined and the query list is modified to contain the IDs and the corresponding number of the coincidence of this ID in the query list.
4. This modified query list is sorted according to the descending number of the coincidences per each ID.

After these steps the IDs of the most similar images will be in the beginning of the query list

The similarity between the query and the images from image database is defined as:

$SIM = \frac{BR_{eq}}{256} * 100\%$ - if the query type is ***Query by Image Example***

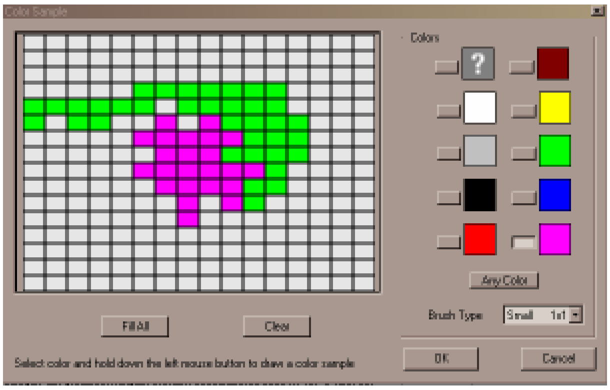
or:

$SIM = \frac{BR_{eq}}{BR_{spec}} * 100\%$ - if the query type is ***Query by User Sketch***,

where:

BR_{eq} - is the number of the blocks in which the dominant color is the same as this in the query, this is the number of the coincidence in the temporal query list for every ID;

BR_{spec} - if the query type is ***Query by User Sketch*** it is possible that the user is looking for similar images not as a whole, but only for images that contain a specific color in a specific position in the image and he does not define the dominant colors in all blocks. If some of the dominant colors are not defined this means that they are not important for the query and they will not take part in the results. (*Figure 2*). In this case the temporal query list contains only the lists for the dominant colors of the specified blocks and only these blocks are analyzed.



Select color and hold down the left mouse button to draw a color sample

Figure 2. ***Query by User Sketch*** *– the user is looking for images that contain purple and green in the center of the image and the rest part of the image is not important.*

CONCLUSIONS AND FUTURE WORK

The main disadvantage of this approach for large image database organization is the number of lists – 2560, which in practice are separate files.

On the other hand, on executing different type of queries is not needed to search all of these large number of lists but only 256 of them or less if the query type is ***Query by User***

Sketch, and the user is specified not all blocks in the query. The number and the exact lists are defined by the query matrix. ***ColorDescriptorMatrix*** of the query defines the exact numbers of the lists using this simple name convention for the files of lists:

- First two numbers in the name of the list are indexes in the matrix (*Figure 1*) and they can be from 1 to 16;
- Third number in the name is defined by the number of the dominant color in this block (table 1 and figure 1) and the possible values are from 0 to 9. This number is equal to the corresponding value in the matrix in the position of two first numbers.

Efficiency of this approach for image database organization and retrieval depends on the selected algorithms and methods for fast file sorting and searching. These algorithms are well known and widely used in the field of theory and practice of the algorithms and data structures.

As a conclusion it is useful to say that the proposed approach is very efficient in respect of similarity calculating, because the similarity coefficients are calculated automatically through the process of list scanning. The number of the blocks with the same dominant colors is equal to the number of the repetition of the corresponding image ID in the final query list. The most similar images are in the beginning of this list.

ACKNOWLEDGMENTS: This work is supported by Bulgarian Ministry of Education and Science – Grant BG 18/05 and in accordance with the Bulgarian-Greek Joint Research Programme.

REFERENCES

1. Boris Rachev, Irena Valova, Silyan Arsov, An Approach for Image Organization and Retrieval in Realistic Image Databases, 7th EC-GIS Workshop, EG II-Managing the Mosaic, Potsdam, Germany, 13-15.06.2001
2. Valova, B. Rachev, Image Databases – an Approach to Image Segmentation&Color Reduction Analysis&Synthesis, International Conference CompSysTech'2003, Sofia, Bulgaria, 19-20 June 2003
3. Valova, B. Rachev, Retrieval By Color Features In Image Databases, Adbis'04, 22-25 September, 2004, Budapest, Hungary
4. Valova, Irena; Rachev, Boris (Bulgaria): "An Algorithm for Organization and Retrieval by Color Features in Image Databases", International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2004; July 21 - 25, 2004 in Orlando, Florida, USA, Volume IV; http://www.infocybernetics.org/citsa2005/pastproceeding/contents.asp?id_volumen=4&action=1&id_conference=6&vpart=
5. Swain, M. J. and Ballard, D. H. (1991). Color indexing. International Journal of Computer Vision, 7(1):11–32.
6. Schettini, R., Ciocca, G., and Zuffi, S. (2000). Color in databases: Indexation and similarity. In Proc. of Int'l Conf. on Color in Graphics and Image Processing, pages 244–249.
7. Schettini, R., Ciocca, G., and Zuffi, S. (2001). Color Imaging Science: Exploiting Digital Media, Ed. R. Luo and L. MacDonald, chapter A Survey on Methods for Colour Image Indexing and Retrieval in Image Database. John Wiley
8. C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. Journal of Intelligent Information Systems, 3(3/4):231–262, July 1994.
9. S. Siggelkow. Feature Histograms for Content-Based Image Retrieval. Ph.D. thesis, University of Freiburg, Institute for Computer Science, Freiburg, Germany, 2002.

10. C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In Int. Conf. Visual Information Systems, pp. 509–516, Amsterdam, The Netherlands, June 1999.
11. D.M. Squire, W. M¨uller, H. M¨uller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In Scandinavian Conference on Image Analysis, pp. 143– 149, Kangerlussuaq, Greenland, June 1999.
12. J.Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries. IEEE Trans. Pattern Analysis and Machine Intelligence, 23(9):947–963, Sept. 2001

ABOUT THE AUTHORS

Irena Valova, University of Rousse, Bulgaria, Phone: ++359 82 888695, Irena@ecs.ru.acad.bg
Boris Rachev, TU – Varna, Bulgaria, Phone: ++359 52 383407, e-mail: Bob_Ra@acm.org
Michael Vassilakopoulos, Technological Educational Institute of Thessaloniki,Thessaloniki, Greece, email: vasilako@it.teithe.gr