A Content Based Image Retrieval System Based on Color Features

Irena Valova, University of Rousse "Angel Kanchev," Department of Computer Systems and Technologies, Rousse, Bulgaria, Irena@ecs.ru.acad.bg
Boris Rachev, Technical University–Varna, Department of Computer Systems and Technologies, Varna, Bulgaria, Bob_Ra@acm.org

Abstract:
Significant research has focused on determining efficient methodologies for retrieving images in large image databases. Most Content Based Image Retrieval systems use low-level visual features for representation and retrieval of images. This paper addresses the design and implementation of a new image abstraction technique based on two compact signatures bit-strings and an appropriate similarity metric. It focuses on a low-dimensional global color features and spatial color distribution based indexing technique for achieving efficient and effective retrieval performance. We propose a combined index structure using these color features. Images are indexed by dominant colors and similar images form an image database cluster stored in a hierarchical structure. The regions within an image are further representing by their dominant colors and this color distribution representation is invariant to translation, rotation and scaling. A query engine supporting tree type of queries (query by image example, query by user sketch and query by global color features) is build in the prototype system to retrieve images by global and local color features. The retrieval performance is studied with a prototype system for content based image organization and retrieval, developed in C++ for Windows and an example collection of 3000 heterogeneous images from www.freefoto.com.

1. Introduction

Retrieval of images by color feature has proved as a great challenge, though there has been considerable research into this topic [4, 5, 6, 7]. Defining a color similarity criteria which corresponds to human visual perception is an unsolved problem. In addition, individual differences in perception of images make the problem even more complicated. This evidence has motivated us for investigating new content based retrieval system. Because of diversity and complexity of the photograph images, it is difficult to capture their visual properties such as shape, structure and complexity.

In this paper we describe the design architecture of our current prototype system. Our goal is creation of an advanced query capability that supports query by image example, by global color distribution and by image sketch. Our prototype system provides a platform for the accommodation of algorithm for automatic feature extraction and automatic image indexing, storage and retrieval by color content.

2. Our work in this field

In the process of our research in the field of image database management systems we examined and analyzed the main functions which must be supported by such system. The classification scheme of these main functions and related subtasks, features and details are shown on figure 1. We made the following conclusions after the analysis of this research:

- ***Image Storage***

Segments represent interesting parts of an image. Many image processing applications first reduce their problem by looking only at the segment or region of interest within an image. This pre-processing can reduce the resource requirements substantially. This leads to the image database requirement of segment representation and segment construction.

Image operations should also work on segments without additional work.

Compression techniques to reduce storage can be used:

- to compress each image in isolation;

- to compress a set of images together, it is possible to compress similar images in clusters or any other structures, or to group images by another criteria and compress them together.

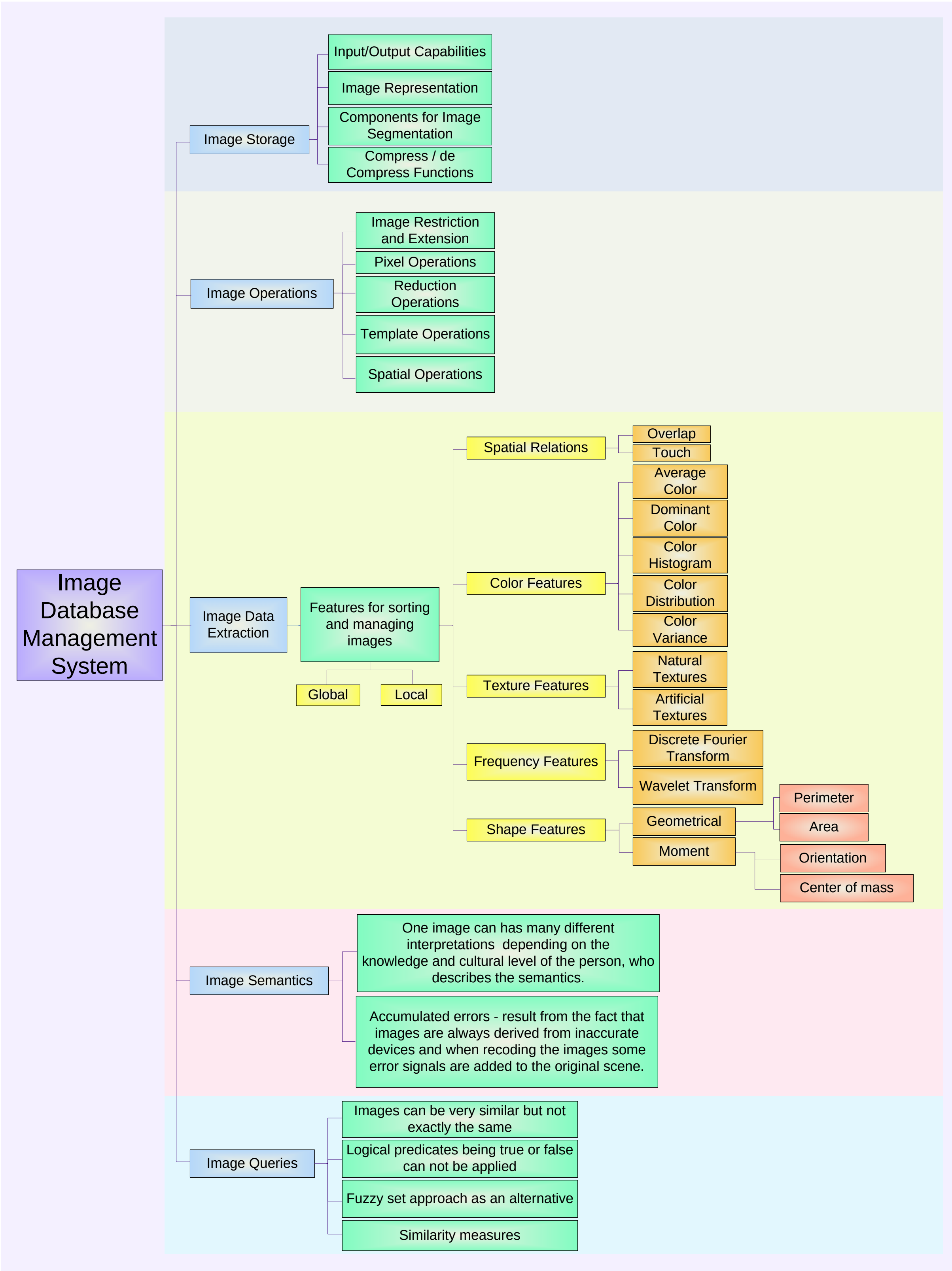


Figure 1. The classification scheme of the main functions and related subtasks, features and details of the image database management systems

- ***Image Operations***

It is necessary for the Image Database Management System to:
- support different types of image file formats;
- support the complete set of image operations

- ***Image Data Extraction***

There are too many image features and the list of features is not complete, but gives a good idea of the various features studied in that field. To be solved specific problems new features are introduced. This means that the Image Database Management System should support an extension mechanism for data structures and operations for representing these features.
Many features are single values (no need for extra data structures), such as area, mean orientation, they can be stored using build-in data types. The representation of the multi value features needs additional data structures.
The local features are calculated over parts of an image – single pixel or segment. Since an image may have several pixels or segments, these features usually result in feature value sets for the whole image. This complicates calculations but it has additional information and can produce better retrieval results.

- ***Image Queries***

In database systems all relational operations are based on logical predicates being either true or false. This rigid logic perspective works, because the semantics of the data entities in business database systems are known and fixed. For image database systems this is no longer sufficient. The world of images is quite fuzzier. Images can be very similar, but are hardly ever exactly the same. This makes it hard to write boolean predicates, such as image equality. A fuzzy set approach is in order here as an alternative.
Image queries are often navigational and managed by a user. Take for example image retrieval systems, which let a user navigate through the image space. In such systems the user constantly refines his query to navigate to the desired image. The reasons for this navigational query approach are two. First, the mentioned fuzzy data makes predicates hard to use. The second reason is that there is still little knowledge of the applications that use an image database. This makes it hard to predict what types of queries are needed, because it is unknown what the interesting data is.
Since the predicate logic expressions are hard to use, they should be replaced by a new technique for comparison. A solution found in many image retrieval systems is based on similarity measures.
The main requirement coming from image queries is a new query model - one possible query model is the fuzzy logic model. Currently, similarity measures are used mainly.

It is easy to understand that the development of such system which covers all of these functions is very difficult task and the most of the existing image retrieval systems realized these functions in part. In our prototype system we realized the following functions:

1. We store the files with images in their format separately from the features derived from the image content. We do not use image compression. We chose to use very simple segmentation approach which divides the whole image into equal non overlapping blocks and it represents the information about the spatial relations between the objects in the image.
   Every image I at pixel level (physical image) is represented in the image database with two type color descriptors and its name and location as:
   $I = (\{\{c_i, p_i\}, i = 1,..., M\}, \{\{C_{i,j}\}, j = 1,.., N, i = 1,.., N\}, \{Path\})$
   This symbolic description of image I is stored in an Image Database.
2. It has been shown that in the early perception stage human visual system performs identification of dominant colors by eliminating fine details and averaging colors with small areas. Consequently, on the global level, humans perceive images only as a combination of few most prominent colors, even though the color histogram of the observed image may consist of

significantly more colors. Because of this we reduce the number of colors to 10 basic colors from our color code book [2, 3].

3. It's possible to use many characteristics during the process of the image content extraction but the creation of a system which works with all of them is difficult. We chose to use the colors as a characteristic of the photograph images content. The system extracts and describes the color content of the images automatically by two structures. First is based on the dominant colors in every image and on the sorted hierarchical tree and second – represents local color distribution. These two structures are stored in the feature database with the path to the image file.
4. In the previous work we made some experiments with the use of key words and annotations to describe the image semantics through the process of adding new image to the image database. The process is very slow and it requires manual work from the user. In the prototype system we do not include these features. In our future work we plan to search a relation between color content end semantics in the images, but this is still difficult and is not a solved problem.
5. The different type of queries, the results from our experiments with this system and the similarity measures used in the system are described in [1, 2, 3]. This is still not well developed and examined and there are still additional problems to be solved and the prototype system to be improved.

3. Architecture of the prototype system

Our prototype system performs two major functions: feature extraction and retrieval. In the feature extraction phase, the system extracts information required for indexing. In the retrieval phase, the information acquired during the feature extraction phase is utilized to assess similarity with a query image.

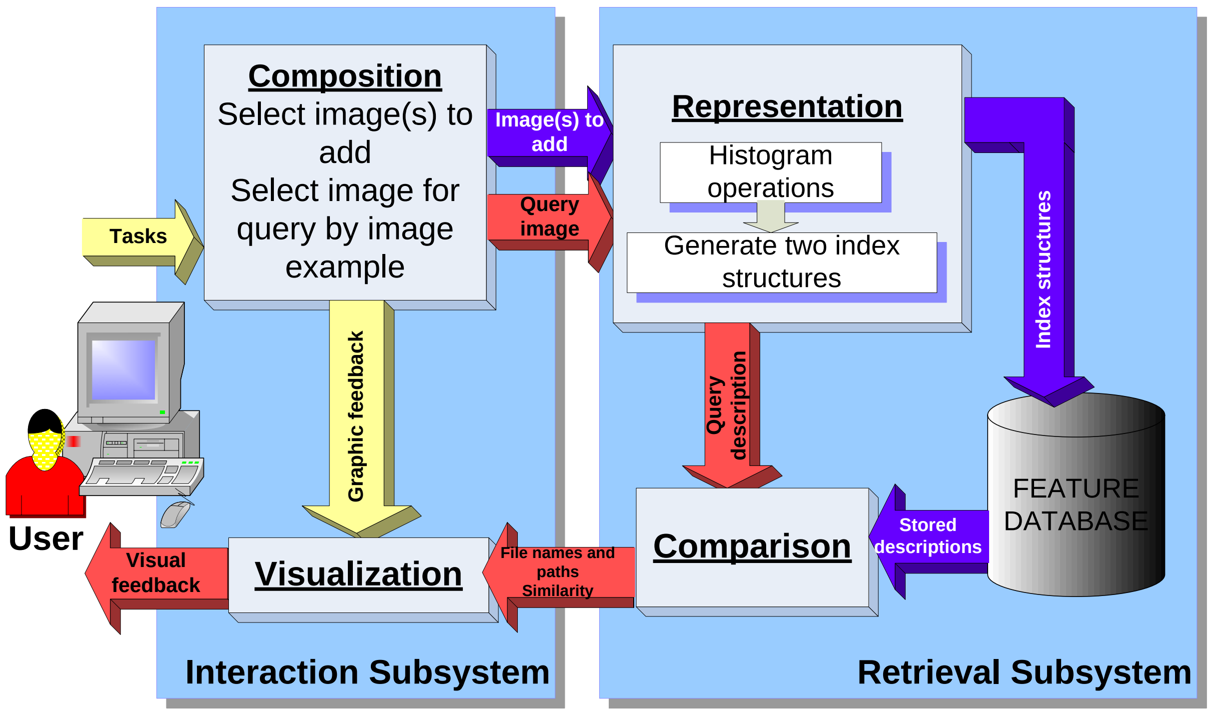


Figure 2. The architecture of the interactive image database management system

As shown in figure 2, 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 the system during the storage process (blue arrows), and become part of the database index. During the retrieval process (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.

The system allows its users the following functionality:

- to create new or open an existing image database
- to add images in an appropriate format to the opened database;
- to build up a database of stored image descriptions from the images in the selected folder;
- to extract retrieval features from these descriptions;
- to allow the formulation of tree different type of visual queries;
- to provide efficient and effective matching of query and stored images;
- to display query results in an appropriate format.
- to browse through the images in the selected database.
- to find duplicate images in the opened image database

The main idea of this approach with ***dominant colors*** [1] is to index the global representation of colors individually rather than form a histogram feature vector and index it in the feature space. It is similar to the multiple text keywords search. Each color can be considered as a keyword. Thus each entry in the database contains several color keywords. During the retrieval process, matching entries containing each color keyword are found and the final results are the join of these matches. In the figure 3 you can see the example of the query by global color distribution and the results from it. On the bases of this representation the sorted hierarchical tree is created. It is described in details in [2, 3].

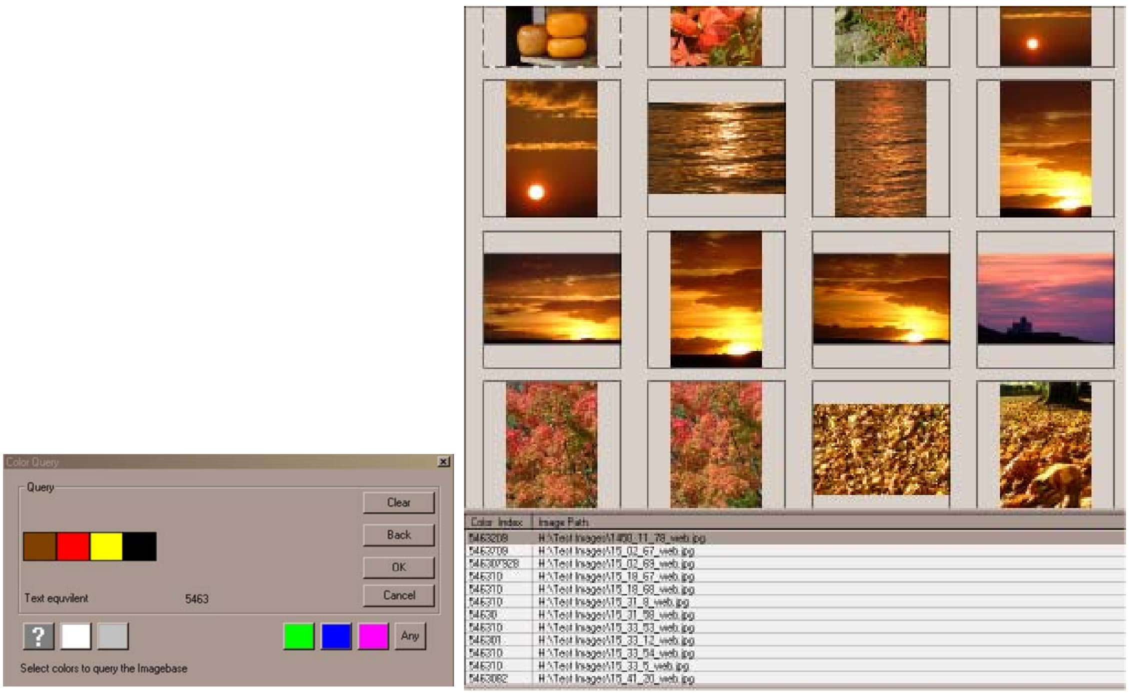


Figure 3. The example of the query by global color distribution (in left) and the results from it (in right). As it can be seen the Color descriptor for all of the returned as a result of this query images begins with the selected in the query colors. (in the first column of the table below the images in right)

2-D vector representation (Color Descriptor Matrix) enables the search for rotated (90, 180 and 270) and horizontally or vertically flipped images, because the direction in which the image is scanned and saved is not known. It is not necessary to save the rotated matrices in the image database. It is enough to search not only for the same matrices but also for rotated ones which is not difficult, as it can be seen from the example. This is done through relatively easy calculations with the positions of the matrix elements. An example of the query by user sketch by local color distribution and its results is shown in figure 4.

The image descriptions mentioned above and used in the prototype system and the algorithms for image organization by color features (OCF) and for **image retrieval by color features (RCF)** are described in [3] and we will not discuss them more here.

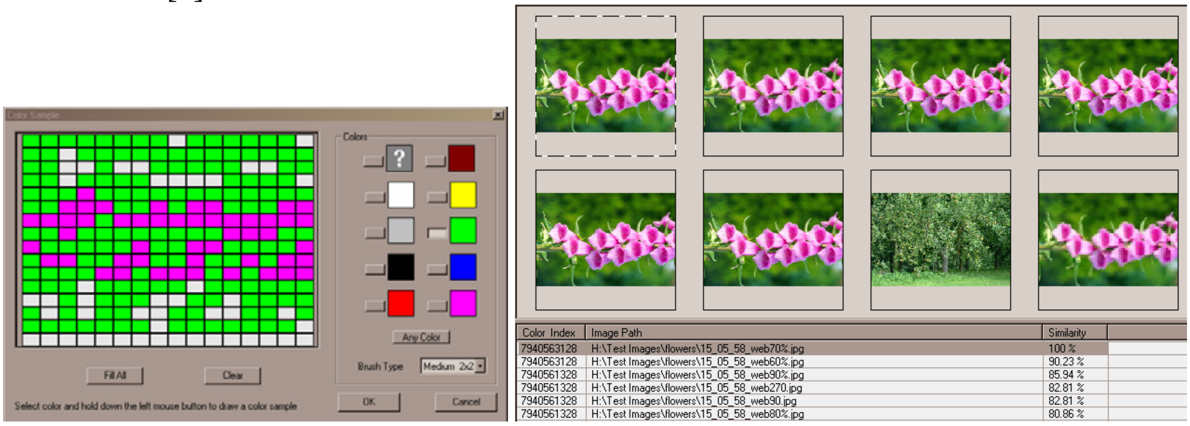


Figure 4. An example of the query by user sketch by local color distribution and its results. There are rotated and scaled images in the results.

Some examples of the organization of the user interface, hierarchical structure and another type of queries supported by the system are shown in the figure 5.

Image browser supports 3 fixed modes for the number of images – 1 image with details, 4 or 16 images without details. The system supports the feature "click to enlarge", which displays the actual size of an image. Also it supports "click to shrink" to return to the display mode of thumbnails. In this window are visualized results from the different type of queries – in this example the query was query by image example and the query image was first left image. The user can select another image for new query or can select another query type or can browse the images from the result set or from the hierarchical tree in the left.

Information Viewer – this is content area to display the related information about all the images retrieved according the selected query type. The output format of the Information Viewer was followed by the common style of text based information systems, which are based on a list results.

Visual representation of the hierarchical tree - user can browse similar images from the different sub-trees by global colors. The initial state of the system is to display all of the images in order of the hierarchical tree and the cursor is located on ALL, which means to display all images. While a specific group is selected by user the system then visualizes only images in this group.

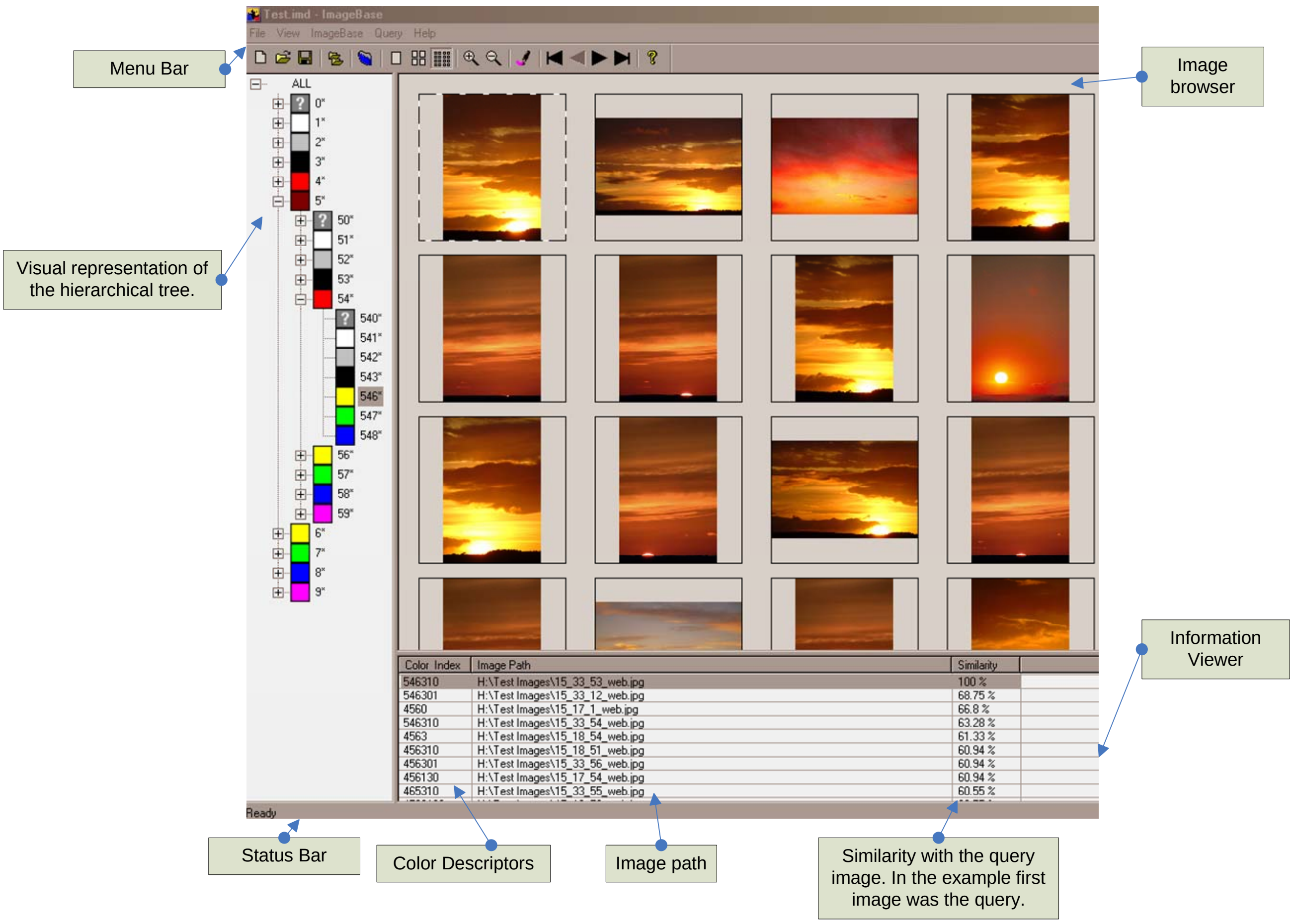
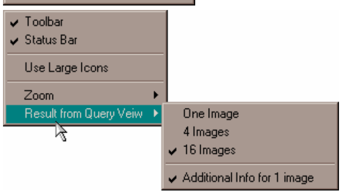


Figure 5. Some examples of the organization of the user interface, hierarchical structure and another type of queries supported by the system.

Menu bar – contains all necessary elements for main functions of the system. They are divided in four different groups – File, View, Image Base and Query.

File – contains functions for creating new image database, adding new images to an existing image database or opening an existing image database.



View – from this menu user can manage how to view the results from the queries, what and how to view from the functions of the system.

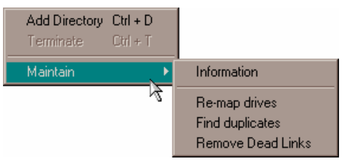
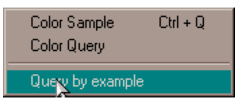
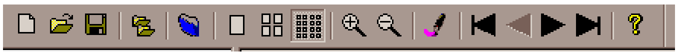


Image Base – the user can manage opened image database.



Query – the query type is specified by this menu.

 **Tool Bar** contains frequently used options from the menus.

4. Conclusions

After the experiments we have made with the system we found that the similarity measures are very important. So we have to improve this in our system. Also we will try to use different well known algorithms for color reduction, not our color perceptual model with 10 pre selected color groups.

5. References

1. Irena Valova, B. Rachev, Image Databases – An Approach for Image segmentation &Color reduction Analysis & Synthesis, International Conference CompSysTech'2003,Sofia, Bulgaria, 19-20 June 2003
2. Irena Valova, B. Rachev, Image organization, querying and retrieval by color distribution features, International Conference CompSysTech'2002, Sofia, Bulgaria, 20-21 June 2002
3. Irena Valova, B. Rachev, Retrieval By Color Features In Image Databases, Adbis'04, 22-25 September, 2004, Budapest, Hungary
4. Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: Past, present and future. Journal of Visual Communication and Image Representation, 10:1–23, 1999.
5. Y. Rui, T.S. Huang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues", Journal of Visual Communication and Image Representation, vol. 10, pp. 1– 23, 1999.
6. A. Goodrum, "Image Information Retrieval: An Overview of Current Research", Special Issue on Information Science Research, vol. 3, no. 2, 2000.
7. Mojsilovic, J. Kovacevic, J. Hu, R. J. Safranek, K. Ganapathy, "Matching and retrieval based on the vocabulary and grammar of color patterns", IEEE Trans. on Image Proc., vol. 9, no. 1, pp. 38-54, January 2000.