A Data Warehousing Approach to Colour Image Retrieval

J. You[1,2] J. Liu[1]
[1] Department of Computing, The Hong Kong Polytechnic University
[2] School of Computing & Information Technology
Griffith University, Brisbane, QLD, Australia 4111

Abstract

This paper describes a new approach to fast content-based color image retrieval in a wavelet-based hierarchical structure with data warehousing techniques. To tackle the key issues such as image data indexing, similarity measures, search methods and query processing in retrieval for large color image archives, we extend the concepts of conventional data warehouse and image database to image data warehouse for effective color indexing. In contrast to the existing systems which employ a fixed mechanism for similarity measurement, we propose to integrate wavelet multi-resolution decomposition with data summarization for hierarchical color representation and flexible similarity measurement. In addition, a guided search scheme is introduced in conjunction with data partitioning and aggregation techniques to speed up query processing. The proposed retrieval method is tested on RGB and YUV color spaces and the experimental results demonstrates its effectiveness and efficiency for content-based color image retrieval.

Key words: content-based image retrieval, color representation and indexing, similarity measurement, guided search, wavelet transforms, data warehouse.

1. Introduction

With the fast growth of multimedia information and wide applications of WWW (world wide web) technology, content-based image retrieval plays an important role in visual information systems. The increased demands to search and browse diverse image databases on the Internet require effective management of image archiving and storage systems. It becomes a very challenging task to store, index, and retrieve huge amounts of image data for scientific and commercial applications.

In general, image retrieval approaches fall into two categories: attribute-based methods and content-based methods [4]. The attribute-based approach is database-oriented by modeling image contents as a set of attributes and managing them within the framework of conventional database-management systems. Such attribute-based methods represent image contents using text and structured fields. The content-based approach is based on the integration of feature-extraction and object-recognition during the management of image databases to overcome the limitation of attribute-based retrieval. Currently a large proportion of research into content-based image retrieval has been centered on issues such as which image features are extracted, the level of abstraction manifested in the features, and the degree of desired domain independence [5].

It is noted that most of the existing techniques for multimedia information retrieval are based on the use of conventional database structures to handle large collections of high-dimensional multimedia data. Although the recent research on multimedia database systems [3], [6], [11] has made advances in creation of large multimedia databases with effective facilities for query processing, it is mainly focused on data modeling and structuring. Object-oriented models have the capacity for retrieving one or more media samples by satisfying a particular set of conditions, however, when faced with large quantities of data that are stored at a low level of information granularity, they can be very slow. Although the application of knowledge discovery techniques to relational multimedia databases has made it possible for multimedia data mining [11], it is difficult to integrate multiple media features for flexible indexing and dynamic search.

Recently the combination of data mining and data warehousing has emerged as an innovative and totally new approach to information management [2]. Data mining provides the capacity for the discovery of hidden knowledge, unexpected patterns and new rules from large databases. A data warehouse is not only a central store of data that has been extracted from operational data, it also contains the process managers that make information available, enabling users to make informed decisions. In other words, data warehousing is concerned with summation, reduction and transformation of data and storing it in a materialized relation available for direct querying. Therefore, it can speed up queries by using a materialized view, and deal with noisy and incomplete data. Consequently, with a data warehouse, we will have clean data, complete data and carry out data reduction and summarization for fast querying with knowledge discovery. Although there have been many successful data mining and data warehousing systems, little has been done on date warehousing and mining for image database for visual information systems.

The main contributions of this paper may be summarized as follows: To tackle the key issues such as image data indexing, similarity measures, search methods and query processing in retrieval for large color image archives, we extend the concepts of a conventional data warehouse and image database to create the notion of an image data warehouse for effective color indexing. In contrast to the existing systems which employ a fixed mechanism for similarity measurement, we propose to integrate wavelet multiresolution decomposition with data summarization for hierarchical color representation and flexible similarity measurement. In addition, a guided search scheme is introduced in conjunction with data partitioning and aggregation techniques to speed up query processing. The proposed retrieval method is tested on RGB and YUV color spaces and the experimental results demonstrates its effectiveness and efficiency for content-based color image retrieval.

This paper is organized as follows. Section 2 reviews the popular algorithms for color representation and introduces a wavelet-based image hierarchy for color content decomposition and multiple color feature extraction. Section 3 outlines the proposed concept of image data warehouse for dynamic color indexing and briefly describes the use of data warehouse schema to integrate multiple color features for hierarchical search and fast query processing. Section 4 describes a coarse-to-fine color content comparison scheme which adopts data partitioning and aggregation techniques to guide the search. Section 5 presents the experimental results. Finally, Section 6 gives our conclusions.

2. Colour Feature Extraction and Representation

Color is an important datum in the real world and has been used as one of the major features for content-based visual information retrieval. The problem of image retrieval by color concerns retrieving all images which have similar color compositions to the query image's color composition. Thus, color indexing is a key issue in relation to the robustness and efficiency of a retrieval system. Many algorithms for color feature representation have been developed for color indexing, and these fall into two categories: global description and local description. This section reviews some of the most popular algorithms and introduces a new color measurement approach based on wavelet transforms.

- Color Histogram

  Color histogram [10] has been adopted by many retrieval systems because of its low memory storage and sufficient accuracy. Although this approach provides a feasible global measurement of color content, it lacks local information about spatial distribution of color composition. For example, the difference between a large region in red color and a large number of scattered pixels in red color will not be captured in a color histogram. To incorporate spatial information into a global color histogram representation, other histogram-based algorithms have been proposed for improvement. in [8].

- Color Moments

  Moments provides a powerful description of geometric properties of an object or a curve distribution such as area, centroid, moment of inertia, skewness and kurtosis. Currently color moments have been adopted for color indexing when an image is represented by its probability distribution function of pixel intensities [9]. It is noted that an image histogram can be represented well by its moments at the first few orders.

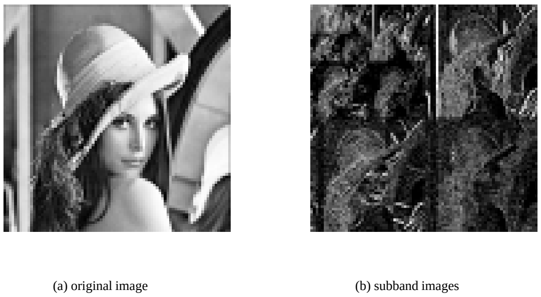
- Wavelet Based Approach

  Wavelet transforms offer the promise of compact representation and efficient detection of image components that match the wave-shape of the chosen wavelet. The examples of wavelet-based coding are summarized in [7], which includes the embedded zerotree wavelet (EZW), the layer zero coding (LZC), set partitioning in hierarchical trees (SPIHT), the rate-distortion optimized wavelet packet (WP). All of these algorithms involve the following three processes: 1) wavelet transformation of a given image; 2) successive approximation quantization (SAQ) of wavelet coefficients; 3) effective coding based on entropy of the resulting coefficients.

- Dynamic Color Indexing – A Proposed Approach

  Color indexing plays a key role in content-based color image retrieval. Indexing tabular data for exact matching or range searches in traditional databases is a well-understood problem, and structures like B-trees provide efficient access mechanisms. However, in the context of similarity matching for visual images, traditional indexing methods may not be appropriate. Consequently, data structures for fast access of high-dimensional features for spatial relationships have to be developed. Nevertheless, the existing approaches are all based on a fixed mechanism and lack an integration of multiple features for dynamic indexing.

With the framework of a color image warehouse which we will introduce in Section 3, we propose a wavelet-based image color hierarchy and multiple feature integration scheme to facilitate dynamic color indexing associated with data summarization. Our approach is characterized as follows: 1) to apply wavelet transforms to decompose a given image into three layers of 10 sub-images; 2) to use the mean of wavelet coefficients at three layers as a global color measurement and index it as tabular data in the global color summary table; 3) to calculate the mean of wavelet coefficients of sub-band images (horizontal, vertical and diagonal) at different layers as local color information and index them as tabular data in a local color summary table; 4) to detect the interesting points of objects in the original image and store them in a table for fine match. To achieve dynamic indexing and flexible similarity measurement, a color data warehouse is introduced in the next section to coordinate data summarization and search for the best match of color similarities. Fig. 1 shows a wavelet-based image decomposition structure for a three-layer image hierarchy.


(a) original image

(b) subband images

Fig. 1. An example of three-level image decomposition

3. Colour Image Warehouse

3.1. An Overview

A data warehouse can be viewed as a technology which not only functions as a data superstore, but also processes data to create a data warehouse, operational data store, or data mart stored on traditional servers, Intranet servers, or Internet servers. In other words, data warehouses are not just large databases; they are large, complex environments that integrate many technologies. Thus, they require considerable maintenance and management. In general, there are four major processes that constitute a data warehouse. which are responsible for the following operations: (1) extract and load the data, (2) clean and transform data into a form that can cope with large data volumes, and provide good query performance, (3) back up and archive data, (4) manage queries, and direct them to the appropriate data sources. It is very important to have the mechanisms that determine when to start extracting the data, run the transformations and consistency checks and so on. Basically, there are two types of tools for the smooth operation of data warehouses – system management tools and data warehouse process management tools.

The data warehouse process managers are software programs responsible for the flow, maintenance and upkeep of the data, both into and out of the data warehouse database. There are three different data warehouse process managers: load manager, warehouse manager and query manager. The load manager is responsible for data source interaction, data transformation and data load. The warehouse manager is responsible for maintaining the data while it is in the data warehouse, which includes data movement, metadata management, performance monitoring and tuning, and data archiving. The query manager is used to control user access to the data, query scheduling and query monitoring. Fig. 2 shows the overall system architecture of a data warehouse.

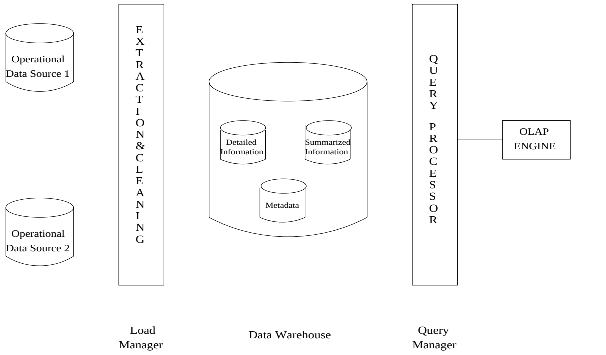


Fig. 2. The overall data warehouse architecture

Developing a data warehouse is similar to other software projects. As summarized in [1], the design process is iterative and consists of the following steps: (1) identify user requirements and project scope; (2) develop subject area data model, including an entity-relationship diagram and associated metadata; (3) develop a data warehouse logical data model from the subject area data model; (4) develop a data warehouse architecture; (5) design the physical database; (6) populate an end-user-oriented repository/directory with metadata for the physical database; (7) identify the source of data from operational systems or external sources for each target field in the data warehouse; (8) develop or purchase programs to extract, cleanse, transform, integrate, and transport data from the legacy systems to the warehouse; (9) populate the warehouse using these programs; (10) test for user satisfaction with the data warehouse, including data quality, availability, and performance; (11) rework the design as needed to achieve ongoing user satisfaction.

3.2. Schema of Colour Warehouse

To achieve the design objectives including performance, future flexibility, scalability of servers, ease of administration, data integrity, data consistency, data availability, user satisfaction, and other factors, an initial critical issue involved is how to design appropriate data warehouse schemas from the logical requirements model. To skip the traditional data modeling process, a star schema is introduced to describe data at a high level.

Schema was initially used in database design. It is a representation of a database, which gives semantics to the data to model the database. In relational databases, there are two types of schema – conceptual schema and logical schema. Conceptual schema often refers to an entity-relationship diagram, and logical schema represents the structure and connection of the relational tables. In object-oriented databases, conceptual models are usually used. Since schema is used at the design stage of any database, it is also adopted to construct data warehouse. Facts and dimensions are two basic concepts related to data warehouse schema. In a conventional data warehouse, facts contain information about an event such as sales, phone calls and bank transactions *etc.*. Fact data is major part of data in data warehouse. It is stable and captured in fact table with multiple foreign keys and millions of rows, where fact table is similar to the relational table in database and foreign keys are used to refer to other tables. Dimensions are used to analyze information, *e.g.* time, location, products and customer groups *etc.*. Dimension data is featured by its changes over time, low volume and different alternatives if necessary. Dimension tables are used to filter out transactions in the fact table that are not required to process a particular query. Therefore, it is very important to identify appropriate facts and dimensions while using schema to describe data warehouse. It is application-dependent to choose specific fact tables for different cases.

Although the concept of data warehouse originated from the traditional business information systems, it can be extended to visual information systems. In this paper, we focused on the establishment of a color image data warehouse for content-based color image retrieval. The content of images is regarded as fact data in image data warehouse, and multiple image features constitute dimension data. With respect to the description of color, mean values of wavelet coefficients of sub-images (global, horizontal, vertical, and diagonal) at different layers, color histogram, color moments and interesting points are considered as dimension data. Fig. 3 shows the relationship between fact data and dimension data, where the center table represents the fact table (content of image) and the surrounding (reference) tables are dimension tables (individual color features). In contrast to the traditional image databases which focus on only image data (content), the proposed image data warehouse extends its capacity for data mapping and summarization. By mapping different image data sources to image data warehouse and generating summary tables based on the dimension data, the retrieval for the similar image by color is guided by directing a query to the right tables for fast searching. The reference to different summary tables will result in the selection of different indexing scheme and similarity measurement. Thus, the use of a data warehouse will facilitate dynamic indexing to speed up the retrieval task.

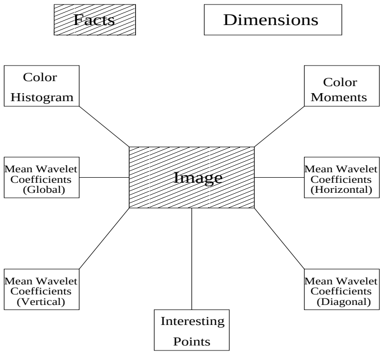

Facts
Dimensions
Color Histogram
Color Moments
Mean Wavelet Coefficients (Global)
Image
Mean Wavelet Coefficients (Horizontal)
Mean Wavelet Coefficients (Vertical)
Mean Wavelet Coefficients (Diagonal)
Interesting Points

Fig. 3. A Data Warehouse Star Schema

4. A Coarse-to-Fine Colour Search

To avoid the blind search for the best fit between the color composition of a query image and the color content of all of the sample images stored in the image data warehouse, a guided search strategy is essential to reduce computation burden. In conventional data warehouse, various methods of partitioning have been developed to improve the efficiency for query processing. The idea behind this is to break up a single entity into multiple ‘smaller’ subentities by slicing the table into smaller pieces. Most of the existing methods fall into two major categories – horizontal partitioning and vertical partitioning. In general, the way in which a fact table will be split up all depends on the type of query. In this paper, we adopted partitioning along a dimension to facilitate a coarse-to-fine color similarity measurement content-based color image retrieval. As stated in Section 3, seven dimensions are associated with a color image – interesting points, global color histogram, color moments, mean values of wavelet coefficients in different directions (global, vertical, horizontal and diagonal).

The initial search for the best similar color composition match starts with the dimension of global mean value of wavelet coefficients. The similarity measurement is based on the difference between the query image and the image samples stored in the image warehouse. The Gaussian normalization procedure is applied before the comparison. The candidates with small distance differences will be considered for fine match in terms of its local wavelet mean values and interesting points. The color histogram and color moments are further used to narrow the selection of the possible candidates for fine matching. The fine match is conducted by using Hausdorff distance of the interesting points associated with similar local color feature.

The color information represented by a star schema in an image data warehouse will facilitate the dynamic search and flexible similarity measurement. The proposed multidimensional structure of the data cube offers flexibility to manipulate the data and view it from different perspectives. Such a structure allows quick data summarization at different levels and the application of OLAP operations (Online Analytical Processing) to view and analyze the data from different aspects as specified. The OLAP operations include drill-down, roll-up, slice and dice. The statistical data resulting from the OLAP operation is used to discover the hidden patterns or implicit knowledge to speed up the task of color composition classification. In other words, we adopted association rules as the technique for data mining to guide the classification.

5. Experimental Results

The color image samples used for the testing are 512×512 size with 256 brightness levels in both RGB and YUV color space. A series of experiments have been carried out to verify the high performance of the proposed algorithms. In our test, a total of 200 image samples are collected. Fig. 4 shows 8 samples of color images selected from the image collection to represent different color content.

Fig. 4. Color image samples

The mean value of wavelet coefficients for the given image at different layers is rotation invariant and can be used as the global index for color content at coarse level, while the mean values of the wavelet coefficients for directional decomposition (horizontal, vertical and diagonal) provide a good description of local color composition which can be used to guide the search at the fine level. Fig. 5 presents a series of rotated images of a given sample and Fig. 6 illustrates the mean value distribution of wavelet coefficients at different orientations.

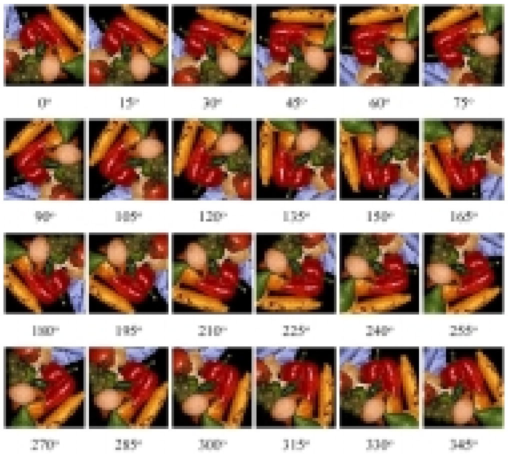


Fig. 5. Sample series of a rotated color image

6. Conclusions

The proposal for an image data warehouse is an effective alternative for fast content-based color image retrieval in a hierarchical manner. The integration of color feature extraction and wavelet transform offers a simple measurement for color content comparison at coarse-level. The use of a subgroup of the color moments and interesting points provides a basis for effective color matching at the fine-level. In addition, the use of data mining techniques on the data cube increases the system flexibility and efficiency. Such an approach will be useful for other multimedia applications.

7. Acknowledgement

The authors are grateful to the partial supports of the University Research Grants (G-YC50 and G-T375) of the Hong Kong Polytechnic University and a Griffith University Research Development Grant for 2001.

References

[1] J. Bischoff and T. Alexander, *Data Warehouse: Practical Advice from the Experts* (Chapter 14), Prentice Hall, Upper Saddle River, New Jersey, 1997.

[2] M.S. Chen, J. Han and P.S. Yu, “Data mining: An overview from a database perspective,” *IEEE Trans. Knowledge and Data Engineering*, vol. 8, pp. 866-883, 1996.

[3] M. Flickner, H. Sawhney, W. Niblack, and J. Ashley, “Query by image and video content: The QBIC system”, *IEEE Computer*, vol. 28, Sept. pp. 23-32, 1995.

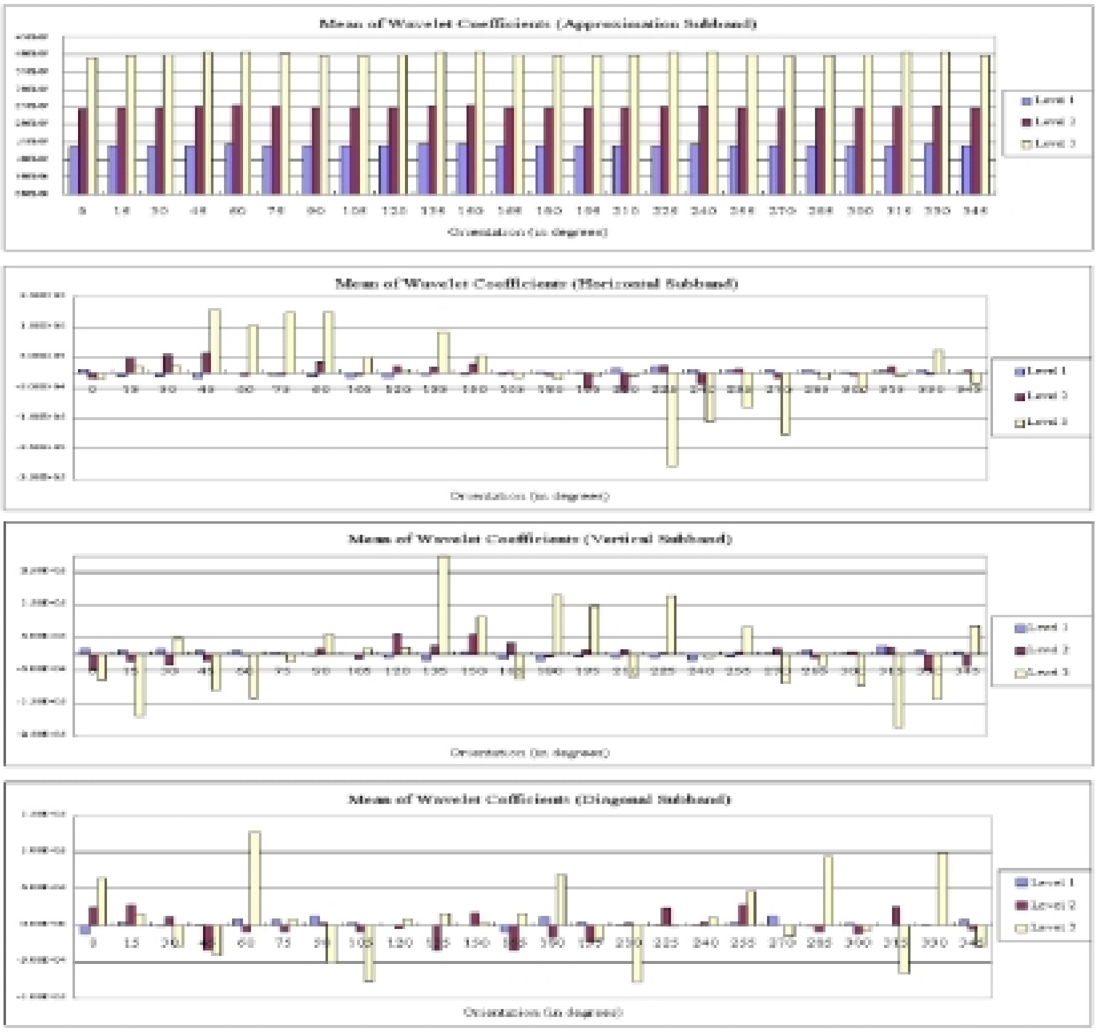


Fig. 6. The distribution of the mean values of wavelet coefficients

[4] W. Grosky and R. Mehrotra, Guest Editors, Special issues on Image Database Management, *Computer*, vol. 22, No.12, Dec. 1989.

[5] V.N. Gudivada and V.V. Raghavan, Guest Editors, Special Issue on Content-Based Image Retrieval Systems, *Computer*, vol. 28, no. 9, Sept. 1995.

[6] S. Khoshafian and A.B. Baker, *Multimedia and Image Databases*, Morgan Kaufmann Publishers, 1996.

[7] K.C. Liang and C.C.J. Kuo, “WaveGuide: A joint wavelet-based image representation and description system,” *IEEE Trans. on Image Processing*, vol. 8, no. 11, pp. 1619-1629, 1999.

[8] W.Y. Ma and H.J. Zhang, “Benchmarking of image features for content-based retrieval,” *Proc. IEEE Conference on Image Processing*, 1998, pp. 253-255.

[9] M.K. Mandal, T. Aboulnasr and S. Panchanathan, “Image indexing using moments and wavelets,” *IEEE Trans. on Consumer Electronics*, vol. 42, no. 3, pp. 557-565, 1996.

[10] M.J. Swain and D.H. Ballard, “Color indexing,” *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11-32, 1991.

[11] O.R. Zaiane, *Resource and Knowledge Discovery from the Internet and Multimedia Repositories*, Ph.D Thesis, Simon Fraser University, 1999.