Image database management systems design considerations algorithms and architecture
Nes, N.J.

Citation for published version (APA):
Nes, N. J. (2000). Image database management systems design considerations algorithms and architecture
Chapter 1

Introduction

Relational Database Management Systems (RDBMS) are commonly used for many business application for almost three decades. Their user friendly declarative query language and abstract logical view on data organization are the main reasons for this success. This declarative query language releases business application developers from dealing with implementation details. They only have to focus on what information is needed, not on how to get it. The later is the task of a query language compiler using a relational algebra as a sound mathematical basis, and gives many ways of optimization.

Since commonly available hardware makes it possible to capture and store images, a need for databases with image management capabilities arose. Large collections of images can easily be obtained from sources like CD-roms, DVD and the Internet. Searching through these collection for particular images is a complex task which requires a proper image retrieval query language. A research problem still lacking general acceptable solutions.

Image databases can be used for many applications. Simple examples include presentations, games and educational software. We could think of an educational program teaching students a foreign language or taking an exam on the traffic rules. In these applications the image database is used as a persistent image store, which gives the user physical data-independence. The image is physically stored somewhere, but the user/application is not interested in its whereabouts. Instead the image can be retrieved using a logical name. Physical data-independence permits a database administrator to move the image without the application noticing it. The advantage is that the images could be scattered over many disks, distributed over various file systems, located worldwide.

Another application domain which would benefit from image database support is image analysis. Image analysis applications try to determine the semantics of an image. The image analysis process involves, segmentation, of images into objects, clustering objects, searching and data reduction. These
operations are strongly supported by database systems. In the image analysis domain each image is a scene taken from the real world. The transition from a real 'continues' world to a 'digital' world introduces many challenging problems. The image analysis domain could benefit substantially from using an image database management system.

Image analysis applications require that images can be accessed on other methods than logical names. Retrieval of stored images requires access methods based on annotations or the image content. Therefore, new query primitives and search methods are needed. A complicating factor is the lack of accurate data. The data involved, both the image itself and the data derived, are fuzzy since perfect image capturing devices do not exist. Also images are subject to varying interpretations. An image will be interpreted different by any other person, caused by the persons background knowledge.

Image database researchers focus on an important subset of applications; called Image Retrieval by Content. These applications try to retrieve images based on the content of the images. A recurring querying scheme is 'query by example'. Given a sample image the system finds a small set of images with similar content. The critical points here are what is considered similar and what content aspects, called features, should be used.

Similarity is often defined as a mathematical function on the features. These so called similarity measures make calculation of similarity possible. Many different similarity measures have been proposed, examples can be found in [63]. Histogram intersection [106] and weighted Euclidean distance [41] are among the most commonly used measures.

To facilitate image queries a two step process is used. First the images are loaded in the database. Secondly for each image feature values are extracted, a fixed set of features is used. The image queries are expressed in terms of these features and selection is based on predefined similarity measures. Most image retrieval systems provide limited control over the features used and definition of similarity measures.

1.1 RDBMS vs IDBMS

Image database management systems (IDBMS) differ from traditional databases management systems in major ways. The first difference is the data complexity. Transaction records are composed of simple data elements like names and numbers. Images, on the other hand, are complex, large arrays of complex values. Also the derived data in an IDBMS is complex, for example at we the contour of an object in an image can be described using a polygon.

On top of the complex data problem is the problem of large object sizes. To illustrate, image databases handle large amounts of sizeable objects (from a few Kilo bytes to several Mega bytes). The granularity of transactions
ranges from a few hundred to a few Kilo bytes. These large data elements result in new requirements for the IDBMS physical storage mechanism.

Query formulation in image databases is more complex. The image domain brings along a large set of operations. The combination of this set of operations inside the realm of relational operators gives an explosion of possibilities making it difficult to oversee. The interplay between the various operators is difficult to predict.

The semantics of the data stored in an IDBMS is unknown. In a traditional transaction processing application all data values are understood by casual users. When we query the database for all persons with age between 18 and 25 we know exactly what we get back. Contrary an image query is worth a thousand words conveying a multitude of semantics. It may well have various interpretations. These semantics have no mathematical basis. This complicates posing and answering queries. A related issue is that data, especially derived data is fuzzy. Many measurement errors may have altered it. The image may largely differ from the real world scene from which it is taken. Therefore, the database should be able to handle incomplete and noisy data.

The last difference from traditional database systems lays in the way the databases are used. The general use of a database in business applications involves many transactional updates, such as inserting, updating and deleting values. In image database systems updates tend to be much more incremental. Once an image is inserted there is little chance that the image is changed or deleted. Therefore, image database applications are much more query intensive than traditional systems. Knowing these characteristics makes it possible to improve the overall system performance.

1.2 Contributions and Thesis Outline

The main contribution of this dissertation is an exploration of facilities needed for the design and construction of a successful image database system. Portions have been implemented in the context of the Monet DBMS to obtain a first assessment of the choices made. However, a complete IDBMS is beyond the scope of this thesis, for it requires a large, multi person, engineering effort. The facilities needed are organized by chapter as follows.

Chapter 2 introduces the basic requirements for an image database, i.e. image storage and management. We founded the requirements on the theory of the image algebra, which supplies us with a complete set of image types and operations.

In Chapter 3 we introduce the physical representation for our image data type. We map the image processing operations on relational operators making it possible for query optimizers to optimize these operations. Also we show possible optimizations of mapping both in storage and processing
requirements.

To obtain the requirements for image retrieval we looked at new image retrieval methods: the multi-level signature and region image indexing.

One requirement obtained using these experiments is the need for a special image retrieval query language. Therefore, we developed the image retrieval algebra which forms the proper basis for such a language. In Chapter 4 this algebra is introduced. Many new database primitives are introduced and their relevance is shown via a preliminary benchmark definition geared at image retrieval queries.

In Chapter 5 we proved the applicability of image databases in the field of image analysis, using a case study. In this study we looked at line clustering, a basic step in many image analyzing applications. We showed databases are effective; they reduce the programming effort and stimulate code reuse, and they are efficient; our implementation proved significant performance improvement over the earlier attempts in this area.

The inherently fuzzy data as found in this field lead to algorithms which search for similar objects. To handle such queries better we introduce in Chapter 6 a new query predicate, the fitness join. Making this key operator explicit at the algebra level provides a handle to optimize its processing. A few directions for optimization are described.

In Chapter 8 we introduce an indexing structure, the metric index structure, to optimize this fitness join operation at a minimum of cost overhead. We showed that metric indexing is a low cost index structure outperforming the R-tree[52]. This structure uses the triangular inequality to reduce the number of calculations of a fitness function.

The Chapters mark a route to a successful IDBMS. The individual steps taken have been commented upon by the research community These Chapters were published earlier in other forms:

- The Multi-level Signature image retrieval method was explained in the paper, 'Database support for image retrieval using spatial-color features' [78].
- An earlier version of the Region Image Indexing was explained in the paper, 'Region-based Indexing in an Image Database'[79].
- In the paper 'Database support for line clustering'[81] image databases were first introduced for image analyzing tasks.
- Metric indexing was explained in the paper 'Metric Indexing to Improve Distance Joins'[82].
- The fitness joins are explained in 'Fitness joins in the Ballroom'[66].
Other related publications not discussed in this thesis:

- The papers, 'Image Retrieval Using Linear Greyscale Granulometries' and 'Color Image Texture Indexing' explain image retrieval techniques where texture is described using object granularity.

In Chapter 8 we summarize the results of the thesis and look forward for further research possibilities.