ABSTRACT
This paper presents the design of a system for content-based image retrieval applied to mammograms. The system takes into account the imprecision present in the search, and makes available user-defined methods to create new image descriptors as well as user-defined feature vectors formed by combinations of previously defined descriptors using a graphical interface. The system was projected and developed using the eXtension Relational DataBase Management System (XRDBMS) PostgreSQL with Image-Handling Extension (PostgreSQL-IE) that supports content-based image retrieval. PostgreSQL-IE is independent of application, and offers the advantage of being open-source and portable. The extended data manipulation and definition language for manipulating image data in the proposed extension is called SQL-IE. The language has a syntax similar to that of SQL (Structured Query Language), and is composed of a set of functions that includes commands to create new feature extraction procedures, new feature vectors as a combination of previously defined features; and new access methods. SQL-IE also includes resources for defining queries by combining conventional and visual data. PostgreSQL-IE makes a new image data type available that permits associating several images with one unique attribute. This resource makes possible the combination of visual features of different images in the same feature vector.

KEY WORDS
content-based image retrieval, feature extractors, Relational Database Management System, mammographic database, image-handling extension

1 Introduction
Content-based image retrieval (CBIR) has emerged as an important area in computer vision and multimedia computing. To support content-based queries with image databases, the descriptors of all images must be accessed, yielding feature vectors to characterize them. The feature vectors are used to retrieve images that satisfy a criterion related to the query. A large number of descriptors may be used to describe an image [1], [2], [3], and feature vectors may be formed as different combinations of the descriptors.

Several works address similarity-based queries for CBIR; however, most of them include methods that are tailored to a particular application [4], [5], [6], [7]. To support application-independent CBIR, commercial extensions are available to Relational Database Management Systems (RDBMS) that are able to store images as a special data type, and to include strategies for searching by similarity restricted to K-Nearest Neighbors (KNN) and range search [8], [9]; examples of such RDBMS are DB2 with Image Extender [10], Informix with Excalibur [11], and Oracle with interMedia [12]. However, these RDBMS possess private architecture, are expensive, and are not flexible to include new feature descriptors or define new feature vectors. Barioni et al. [13] proposed to add new constructions to the SQL language to represent similarity queries that includes KNN, Range, Range Join and Next Join operations. The extension is interpreted by the SIREN (Similarity Retrieval ENgine) engine, that acts as a blade between Oracle 9i DMBS and the application program. The SIREN engine was implemented in C++ under the Windows operating system and does not present flexibility to include new feature extractors or new access methods.

Aiming to have an application-independent RDBMS to support CBIR without extra cost, to make use of the powerful tools available with the traditional RDBMS with free source, to include more flexibility with respect to the creation of new feature descriptors, and to define new feature vectors, we propose an extension to the architecture of PostgreSQL called PostgreSQL with Image-Handling Extension (PostgreSQL-IE). The extension is composed of the system catalog of PostgreSQL with a set of new tables to handle image data, and a packet of functions that maintain the original functionality of PostgreSQL [14], [15], but add special characteristics to deal with CBIR.

To illustrate the power of the resources available in PostgreSQL-IE, we present in this paper a system called SISPRIM — Sistema de Pesquisa para Recuperação de Im-
agencies Mammográficas, that is, a Research System for Retrieval of Mammographic Images — a research system that supports content-based retrieval of mammograms. The research system allows the user to retrieve information from the mammographic database by combining in the condition clause of the query conventional and image data through a friendly graphical Web interface. The proposed research system makes available two similarity search procedures: KNN and range; and seven feature extractors: Fourier descriptors, Area, Perimeter, Compactness, Spiculation Index, Concavity Index [16], [17], and Histogram. The feature extraction procedures are developed in the C language and available in the dll (dynamically linked library) format for the Windows and in the so (shared object) library format for the Linux operating system. Experimental results are shown for validating the system.

This paper is organized as follows. Section 2 presents the research system SISPRIM. Section 3 presents an overview of PostgreSQL-IE, Section 4 presents a discussion of the results, and Section 5 concludes the work.

2 SISPRIM — A System Design for CBIR Applied to Mammograms

This section presents SISPRIM, a research system that supports CBIR, applied to the analysis of mammograms for computer-aided diagnosis (CAD) of breast cancer. The system is composed of a mammographic database modeled using PostgreSQL-IE that includes facilities to organize visual and conventional data, and to answer queries based on similarity in a simple, efficient, and fast manner; a research engine developed in PHP, and a Web Graphical User Interface (GUI).

The mammographic database stores, for each patient, her historical clinical information, some relevant aspects of her life style, and the mammographic exams performed at different instants of time. For each mammographic exam, the following are stored: the four standard views (two views of each breast: cranio-caudal or CC and medio-lateral oblique or MLO); complementary exams; information about the presence of abnormalities, such as architectural distortion, asymmetric density, palpable lump, and calcifications; and the diagnosis, as applicable, according to the BI-RADS classification system [18]. The mammographic database associates, with each mammographic exam, the report given by the radiologist, and links each mammogram of each exam with the contour of the breast, the boundary of the pectoral muscle (MLO views only), the contours of masses (if present), the regions of clusters of calcifications (if present), the number of calcifications (if present), and the locations and details of any other features of interest. The contours of masses and regions of clusters of calcifications may be drawn and entered into the system by an authorized expert radiologist.

The research engine is the heart of the research system: it manages the actions of the user and accesses the mammographic database as required to answer queries. The proposed research system can answer queries as return five images that are similar to the given reference image and the patient takes antidepressive medication, or return the identification of the patient and the diagnosis associated with each one of ten images similar to the reference image, and the density of parenchyma is P2, and the age at menarche is 10 years.

To handle queries as above, the engine system requires that each mammogram in the database be associated with a set of feature descriptors combined into one or more feature vectors, and that the data manipulation language possess resources to combine visual and conventional information in the condition clause of the query. If the RDBMS does not provide support for CBIR, these requirements must be developed in the application system. To overcome this problem, we propose to use the eXtended Relational DataBase Management System (XRDBMS) PostgreSQL-IE. As the mammographic database used by SISPRIM was modeled using PostgreSQL-IE, all requirements for supporting CBIR are available. To facilitate easy access to the information stored in the mammographic database and to access the resources available in the XRDBMS, the research engine manages a Web GUI that incorporates facilities for supporting two kinds of users: administrators and researchers. The interface for an administrator guides the user in the creation of a new feature extractor and/or a new feature vector, and in the association of a given feature vector with an image attribute, including an access method (optional). The interface for a researcher, shown in Figure 1, guides the user through the resources previously configured by an administrator. A researcher can interactively configure a query, combining conventional and visual data, as desired.

3 The PostgreSQL with Image-handling Extension

In a conventional relational database, images are stored as Binary Large OBjects (BLOBs), allowing their retrieval only through textual or numerical keys. To overcome this limitation, some commercial database systems included new data types to represent image data and resources to support CBIR. In general, all of the currently available commercial systems support an image data type that is able to store only one image per image attribute; such a data type encapsulates information about the storage image with some specific differences. The commercial systems present differences with respect to the number of feature descriptors available in the system, the way of managing them, as well as the composition of the feature vectors.

For example, DB2 with Image Extender is limited to four feature descriptors, three of them based on color and the other one based on texture [10]. Oracle with InterMedia [12] is limited also to four feature descriptors, two of them related to color, one related to texture, and one related to
shape. Informix with Excalibur makes available five feature descriptors: two based on color, two based on texture, and one based on shape [11]. In all of these systems, it is possible to associate only one feature vector with each image attribute, and the feature vector has to be composed of the feature descriptors available in the system: no user configuration is permitted.

To develop CBIR systems for medical applications, it is desirable that the RDBMS makes available more flexibility to store the images, to define new feature descriptors, and to combine the features into vectors in new manners that are not predefined. To address these requirements, we propose PostgreSQL-IE that supports a new image data type called PGImage, with the capability to store more than one image in the same image attribute. This facility is helpful in developing applications where each case is composed of a set of images such as computed tomography, magnetic resonance imaging, or mammography. Feature vectors can be associated with a set of images stored in the same attribute, in a flexible way. This resource makes CBIR powerful.

PostgreSQL-IE is configured by adding to the system catalog (Template1) of PostgreSQL an extended system catalog called Extended_Template1. The extended system catalog is composed of a packet of functions written in PLPGSQL [19] and the C language that extend the standard SQL, and six new tables for managing the extension. SQL-IE is composed of a set of sixteen functions that includes commands to create new feature extraction procedures, new feature vectors as a combination of previously defined features, and new access methods. A list of the main functions available in SQL-IE is presented below:

- **Create_Extractor** – This function makes available a new feature extractor by inserting specific information about the feature extractor into the extended system catalog table called pge_extractor, and by creating internal control functions that are able to access the appropriate extractor function library.

- **Define_Feature_Vector** – This function defines a new feature vector as a combination of data. The extended catalog system tables pge_vector and pge_vector_extractor are updated.

- **Create_AccessMethod** – This function creates a new access method. The table pge_accessmethod of the extended catalog system is updated.

- **Set_Feature_Vector** – This function associates an image attribute of a given database table with a previously defined feature vector, and creates an index structure if an access method is defined. More than one feature vector can be associated with the same image attribute in the same table. In order to avoid allocating extra storage space and to avoid repeated calls to execute the same function, PostgreSQL-IE reuses the feature extractors that have already been created.

The table pge_tabvector in the extended system catalog relates columns in the internal control tables to the respective feature vectors.

- **Insert_Image** – This function is used as part of the conventional SQL insert command. The Insert_Image function allows the user to store one or more images in the same image attribute, and to launch the execution of all the feature extractors of all the feature vectors associated with the image attribute being inserted. Additional information such as the width and height of the images are also saved.

Other functions include Delete_Extractor, Width_IE, Delete_Feature_Vector, Delete_AccessMethod, Show_Image, UnSet_Feature_Vector, Show_Extractor, Update_image_Attr, Replace_Image_Attr, Height_IE, and Value_Extractor. To create a new CBIR application system using PostgreSQL-IE, it is just necessary to use Extended_Template1 as the template of the new database application.

PostgreSQL-IE does not yet have any multidimensional access method available. The queries are carried out by scanning all the tuples in the specified table. This procedure is referred to in this paper as Standard-Access-Method, and includes KNN search and RANGE search. Although there is no query optimization, the similarity operators allow CBIR using a feature vector defined as a combination of several features with different weights. The distance between two images is obtained by computing a weighted sum of the normalized difference of each feature value. For carrying out KNN or RANGE search, the differences between a given reference image and all the images in the referenced database table are computed, the images are ranked in increasing order of the resulting distances, and the similarity operations are appropriately processed. The proposed extension is portable, easy to be installed, and is available for downloading at www.lcc.ufu.br/pdi/download.

### 4 Discussion of the Results

This section presents an example to illustrate the functionality of SISPRIM. The dataset of contours and ROIs (Regions of Interest) of breast masses used in this example includes images obtained in two preceding studies. One set of images was derived from mammograms of 20 cases obtained from Screen Test: the Alberta Program for the Early Detection of Breast Cancer [20]. The mammograms were digitized using the Lumiscan 85 scanner at a resolution of 50 µm with 12 b/ pixel. The set includes 57 ROIs, of which 37 are related to benign masses and 20 are related to malignant tumors. Another set of images was obtained from the Mammographic Image Analysis Society (MIAS, UK) database [21] and the teaching library of the Foothills Hospital (Calgary) [16]. The MIAS images were digitized at a resolution of 50 µm; the Foothills Hospital images...
were digitized at a resolution of 62 µm. This set includes smooth, lobulated, and spiculated contours in both the benign (28) and malignant (26) categories. The contour of each lesion was manually drawn by an expert radiologist specialized in mammography. The combined dataset has 111 contours, including both typical and atypical shapes of benign masses (65) and malignant tumors (46).

PostgreSQL-IE has seven feature extractors available for the present example, the administrator defined the feature vector called vector1 composed of Spiculation Index, Compactness, and Histogram, and set vector1 to the attribute mammo, in the table Cases of the mammographic database, using the access method Standard-Access-Method, as described in Section 3. This setting is called Condition1.

Figure 1 illustrates the definition of the query return three images similar to the given reference image, with the condition that the patient had her first pregnancy after the age of 40 years. The answer includes the three rank-order similar images, the id of the patient, and the age of the first pregnancy of the patient. At the bottom of the interface in Figure 1, the SQL-IE command that represents the defined query is displayed. Figure 2 shows the result of the executed query. Note that the reference image does not appear as part of the answer to the query because the age of the first pregnancy of the patient is less than 40 years, which does not satisfy the defined condition. (Fictitious ages have been used in this example).

5 Conclusion

We have presented two important contributions. The first one is PostgreSQL with Image-handling Extension, an XRDDBMS that supports CBIR in a flexible way. The system is open-source, portable, easy to be installed, and is available via the Web for the Windows and the Linux operating systems. At present, the system includes seven feature descriptors (Compactness, Area, Perimeter, Fourier Descriptors, Fractional Concavity Index, Spiculation Index, and Histogram). Further work is being conducted to make available three more shape descriptors: Fractal Dimension, Spiculation Index based on the Turning Function, and Convexity Index [22], [23]. PostgreSQL-IE supports a new image attribute type that permits the user to model a relational schema by storing various images of different classes in the same attribute. This resource makes possible combining feature extractors of different images in the same feature vector. To date, PostgreSQL-IE makes available two conventional similarity search procedures (KNN and RANGE), but new studies are being conducted to extend PostgreSQL-IE with two new similarity search procedure based on fuzzy sets to take into account the uncertainty present in CBIR.

The second contribution is SISPRIM, a research system designed by using PostgreSQL-IE. The research system accesses a mammographic database that includes complete information about the patient (clinical history and life style), her mammographic exams, complementary exams, and the report of each exam associated with the radiological findings of each view of each mammogram, according to the BI-RADS classification system. The information about the clinical history and life style includes details such as the use of alcohol, tobacco, antidepressive medication, hormone replacement; age at menarche; age at first pregnancy; and menopausal status. The radiological findings include asymmetric density, architectural distortion, parenchyma type, and descriptions of the characteristics of masses and calcifications (as applicable). With information as above, it is possible to compute statistical measures and correlation coefficients to derive a relationship between the incidence of breast cancer and the life style of the patient. The system also permits temporal analysis of the evolution of the breast, which is useful in understanding the natural changes of the breast and interval cancer. SISPRIM makes available a graphical Web interface that helps the user to configure and execute queries. To illustrate the power and facilities of SISPRIM and PostgreSQL-IE, we have modeled a mammographic database with 111 images and carried out queries using the KNN operator. The relevance of the results depends on the feature extractors and the feature vectors used. Using a dataset of images of 57 breast masses, Alto et al. [1] obtained a precision of retrieval of 95% using only the shape factor of fractional concavity. Tests with a new dataset are in progress to evaluate the retrieval performance several combinations of image-based and clinical features.


Acknowledgement

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil and the Catalyst Program of Research Services, University of Calgary, Canada.

References


Figure 1. The researcher graphical Web interface for defining a query.
Figure 2. The results of the search as defined in Figure 1.