

# MammoSVD: a Content-Based Image Retrieval System Using a Reference Database of Mammographies

Júlia E. E. de Oliveira, Ana Paula B. Lopes, Guillermo Cámara-Chavez, Arnaldo de A. Araújo  
Federal University of Minas Gerais - Department of Computer Science  
Av. Antônio Carlos, 6627 - Belo Horizonte, MG, Brazil  
{julia, paula, gcamarac, arnaldo}@dcc.ufmg.br

Thomas M. Deserno  
Aachen University of Technology - Department of Medical Informatics  
Pauwelstrasse 30 - Aachen, Germany  
tdeserno@mi.rwth-aachen.de

## Abstract

*In this paper, we present a content-based image retrieval (CBIR) system called MammoSVD. This CBIR system is developed based on breast density – fatty or dense, and the database used, from the IRMA project, provides images with the ground truth already set. Singular value decomposition (SVD) is proposed for the breast density characterization by the selection of the first singular values, in order to represent texture along with the dimensionality reduction. Support-vector machine (SVM) is used to perform the retrieval operation. Considering the first 10% of the retrieved images, the precision rate is 90%, indicating the potential of the implemented CBIR system.*

## 1. Introduction

Medical images are important for diagnosis purposes as they are related to patient's medical historic and pathology. Mammography uses low x-ray doses to produce images of breasts and it is an efficient and largely used method to the early detection of breast cancer. Breast cancer represents one of the main causes of death among women in occidental countries (Brazilian National Cancer Institute, <http://www.inca.gov.br>).

Breast density has been shown to be related with the risk of the development of breast cancer [19] since women with a dense breast density can hide lesions and so cancer is detected at later stages. A density scale named BI-RADS (Breast Imaging Reporting Data System) developed by the American College of Radiology (<http://www.acr.org>) informs radiologists about the decline in sensitivity of ma-

mmography with increasing breast density. BI-RADS defines density 1 as almost entirely fatty, density 2 as scattered fibroglandular tissue, density 3 as heterogeneously dense tissue and density 4 as extremely dense tissue.

Radiologists evaluate and report breast density on the basis of the visual analysis of mammographies. Computer-aided diagnosis (CAD) and content-based retrieval (CBIR) systems appear as a real possibility to help radiologists in reducing the variability of their analysis. CBIR systems use visual information extracted from images to retrieve similar images to one query image and this system does not need to provide diagnosis information of the retrieved images but just present similar images according to a certain pattern. Considering a CBIR system based on the breast density, from a clinical point of view, such a system can guide the radiologist for the detection of a lesion and its classification. Moreover, from a technical point of view, this system is the first step, and a very important one, for the development of a CAD system.

In this work, we propose, implement, and evaluate a CBIR system called MammoSVD. The breast density is characterized through singular value decomposition (SVD) [5], and the support vector machine (SVM) [16] classifier is used for the retrieval task.

In the context of mammography and breast density, some works explored the use of CBIR and CAD systems. Kinoshita *et al* [8] used breast density as a pattern to retrieve 1,080 mammographies from the Clinical Hospital from the University of São Paulo, Ribeirão Preto, Brazil. Shape descriptors, texture features and histograms were used to characterize the breast density, and the Kohonen self-organizing map (SOM) neural network was used for the retrieval task. Precision rates between 83% and 79%

were obtained for 50% and 25% of recall. Despite the fact that these results indicate that through certain types of features, such as histograms and shape, retrieval concerning the breast density can be effective, and additional studies are needed to improve all the process of retrieval.

Regarding only the breast density classification, mammographies were automatically divided between fatty tissue and dense tissue [13] using the Fuzzy-C means algorithm. From all images, texture and morphological features were extracted and then mammographies were classified using decision trees and k-Nearest Neighbor algorithm. Comparing with radiologists' classification, the results are of 86% of correct classification for MIAS (The Mammographic Image Analysis Society Digital Mammogram Database) database [15] and 73% for DDSM (The Digital Database for Screening Mammography) database [7].

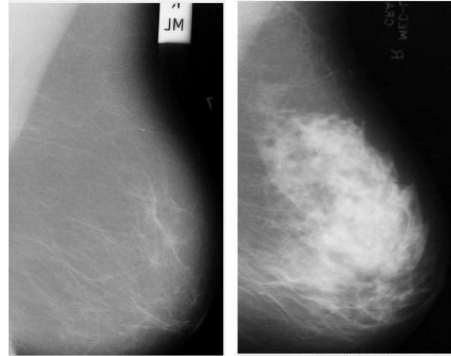
A main challenge for the development of CBIR systems is the appropriate characterization of images and the storage and management of the big amount of images produced by hospitals and medical centers. The IRMA (Image Retrieval in Medical Applications) project deals with this kind of problems, as it aims at developing and implementing high-level methods for CBIR systems with prototypal application to medico-diagnostic tasks on radiological image archive [11]. There are currently more than 30,000 diagnostic images with available ground truth information in the IRMA database. They are used for image retrieval and computer-aided diagnosis [3, 10]. Regarding mammography, there are more than 10,000 images in the database [2], all of them also with available ground truth information. This database offers invaluable support to the validation of the method proposed in this work.

The remainder of this paper is broken into five sections. Section 2 introduces the texture characterization of the breasts through SVD. In Section 3, we expose the basic principles of the SVM classifier used for the retrieval task. Section 4 presents the experiments. In Section 5, we present and discuss the results, and in Section 6, we state the conclusion of the work.

## 2. Breast Density Characterization

In CBIR systems, the access to information is performed by the visual attributes extracted from images. The definition of a set of features, capable to describe effectively each region contained in an image, is one of the most complex tasks in the analysis of images. In addition, the process of characterization affects all the subsequent process of a CBIR system [1].

An image can be numerically represented by a feature vector, which should reduce the dimensionality of the image and emphasize aspects of this image [4]. Visually, breasts



**Figure 1. On the left, mammography of fatty density. On the right, mammography of dense density.**

of fatty and dense densities differ through gray level intensity in mammographies, as can be seen in Figure 1. Since texture contains information about the spatial distribution of gray levels and variations in brightness, its use for the representation of breast density becomes appropriate [6].

The high dimensionality of the feature vector is one of the difficulties in the use of the texture attribute, so it is desirable to choose a technique that combines the representation of this texture with the reduction of dimensionality, in a way to turn the retrieval algorithm more effective and computationally treatable.

The method of SVD consists in decomposing a matrix, whose elements can be composed of the intensity of the pixels belonging to a certain texture, in a matrix multiplication operation [9, 17]. The singular values obtained as results of this decomposition provide useful information of the texture, and for purposes of reduction of dimensionality only the first  $k$  singular values are kept. The goal is to find the best rank  $k$  that would improve the image characterization [9], and this rank  $k$  must be no more than the minimum value between the sizes of the matrix.

## 3. Support Vector Machine for Content Based Image Retrieval

Image retrieval has the purpose to retrieve, from a database, images that are relevant for one query. The query image goes through the process of extraction of attributes and the obtained feature vector is submitted to a search for similarity together with the structure containing the feature vector of all images stored in the database. The identities of the resulting images from the search are used to retrieve these images from the database, thus they can be presented to the radiologist.

MammoSVD deals with a binary classification of data:

fatty breast density or dense breast density. The support vector machine (SVM) method is considered a good classifier since it is able to predict correctly the class of the new data from the same domain where the learning occurred [14, 18].

SVM can be described for a binary classification as follows: given two classes and a set of points that belong to these classes, the SVM classifier determines the hyperplane in the feature space that separates the points in order to place the highest number of points of the same class on the same side, while maximizing the distance of each class to that hyperplane. The hyperplane generated is determined by a subset of items from the two classes, called support vectors.

When the sets of data are linearly separable by a straight line, it is called a linear case of separation. But in most of the cases, this linear case is a restrictive hypothesis to be used in practice. So, instead of a straight line, it is used one function called kernel. The most commonly used kernels are the polynomial and Gaussian ones [16].

## 4. Experiments

The MammoSVD system uses mammographies from the database of radiological images from the IRMA project [2]. In the IRMA project, all images are coded according to a mono-hierarchical, multi-axial coding scheme [12], and this codification provides the ground truth of all mammographies, as all the images were previously verified by an experienced radiologist.

The images, which have approximately  $1,024 \times 500$  pixels of size and are from both medio-lateral and cranio-caudal projections, were grouped in mammographies of fatty density – 200 mammographies from ACR BI-RADS 1 and 200 mammographies from ACR BI-RADS 2 – and mammographies of dense density – 200 mammographies from ACR BI-RADS 3 and 200 mammographies from ACR BI-RADS 4.

For all the images, in a way to remove noises such as black areas and exams labels, it was performed a segmentation of the breast region.

After this segmentation, the steps followed for the development of the CBIR system were:

**Step1** → **Extraction of singular values:** the following first  $k$  singular values were kept for the composition of the feature vector: 25, 50, 75, 100, 150, 200 and 250. These values were chosen empirically according to [9].

**Step2** → **Measurement of similarity between images:** SVM computes the similarity between images through the indication of relevance of the image to a certain query. The set of 800 feature vectors was used in the following manner:

- Training: 240 feature vectors of fatty breast density and 240 feature vectors of dense breast density.

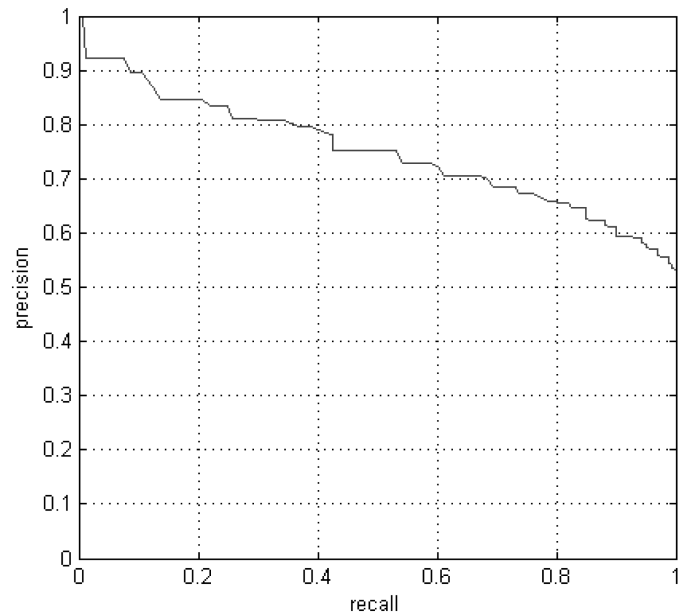
- Test: 160 feature vectors of fatty breast density and 160 feature vectors of dense breast density.

The selection of the feature vectors used for training and the ones used for test was done randomly, and the two sets are disjointed. Moreover, tests were done using the linear case and the polynomial and Gaussian kernels. Of the three cases, the polynomial kernel was the one capable of separating more efficiently the two classes.

**Step3** → **Evaluation of the CBIR system:** measures of precision and recall were obtained and all the 320 mammographies that were not used for the training of the SVM classifier were used as query. We considered values of precision for 10% of recall since radiologists pay more attention to the top returned images.

## 5. Results and Discussion

Figure 2 presents the precision and recall curve for the best result obtained, the one using the first 200 singular values for breast density characterization and SVM for image retrieval. This selection allows representing the texture of the mammographies together with the reduction of dimensionality, as storing only few values results in significant computational savings over storing the whole vector.



**Figure 2. Precision and recall curve for the first 200 singular values for breast density characterization.**

A value of 90% of precision for 10% of recall means that from 32 mammographies returned by the MammoSVD

system, 29 images are relevant for the query user. The SVD method was able to capture the difference between the gray level intensities of the breast densities and characterize them. Also, because of the high generalization ability of the SVM classifier, the training with the polynomial kernel was able to separate the two classes and indicate the most relevant images to the query one.

Although radiologists look for breast lesions like masses and calcifications in mammographies, a CBIR system for mammography should include all possibilities. MammoSVD is the first stage of a CBIR system, as the breast density plays an important role in the diagnostic process.

An important characteristic of the proposed CBIR system is the use of a *priori* breast density classification, as all the images contained in the IRMA database have their ground truth already set by an experienced radiologist.

Future works will focus on the retrieval of four classes of breast density according to ACR BI-RADS scale and on the visual presentation of the retrieved images.

## 6. Conclusion

The research on CBIR still has some challenges. One is which approach to use, more efficiently, to characterize images through a small sequence of numerical values, thus reducing the dimensionality of the original images and how to represent these features properly. In this paper we presented a CBIR system, called MammoSVD, which uses the breast density as standard for image retrieval as this can hide lesions indicative of breast cancer. The mammographies used, belonging to IRMA database, are already classified, setting the ground truth, in order to provide for the retrieval process, beyond similar images, also diagnosis information of the mammographies.

In this system, the segmented areas of the breast are characterized through SVD and only the first singular values are kept, in a way to represent texture together with the reduction of dimensionality of the feature vector. This characterization allied with SVM classification for image retrieval enables the development of a CBIR system of mammographies that can really aid radiologists in their diagnosis.

## Acknowledgment

This work is supported by CAPES and CNPq, Brazilian research funding agencies. The IRMA project is funded by the German Research Foundation (DFG), Le 1108/4 and Le 1108/9.

## References

[1] R. Baeza-Yates and B. R. Neto. *Modern Information Retrieval*. Addison-Wesley Professional, 1999.

[2] J. E. E. de Oliveira, M. Güld, A. de Albuquerque Araújo, B. Ott, and T. Deserno. Towards a standard reference database for computer-aided mammography. In *Proceedings of SPIE Medical Imaging*, volume 6915, page 69151Y, USA, 2008.

[3] T. Deselaers, H. Müller, P. Clough, H. Ney, and T. Lehmann. The CLEF 2005 automatic medical annotation task. *International Journal of Computer Vision*, 74(1):51–58, 2007.

[4] R. O. Dudda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley Sons, 2001.

[5] G. H. Golub. *Matrix computations*. Johns Hopkins series in the mathematical sciences, 1983.

[6] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital Image Processing using Matlab*. Prentice-Hall, 2003.

[7] M. Heath, K. Bowyer, and D. K. et al. Current status of the digital database for screening mammography. In: *Digital Mammography*, Kluwer Academic Publishers, pages 457–460, 1998.

[8] S. K. Kinoshita, P. M. de Azevedo Marques, R. R. P. Jr, J. A. H. Rodrigues, and R. M. Rangayyan. Content-based retrieval of mammograms using visual features related to breast density patterns. *Journal of Digital Imaging*, 20(2):172–190, 2007.

[9] L. Elden. Numerical linear algebra in data mining. *Acta Numerica*, pages 327–384, 2006.

[10] T. M. Lehmann, M. O. Güld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and B. Wein. Automatic categorization of medical images for content-based image retrieval and data mining. *Computerized Medical Imaging and Graphics*, 29(2):143–155, 2005.

[11] T. M. Lehmann, M. O. Güld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. Wein. Content-based image retrieval in medical applications. *Methods of Information in Medicine*, 43(4):354–361, 2004.

[12] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. Wein. The IRMA code for unique classification of medical images. In *Proceedings of SPIE*, volume 5033, pages 440–451, 2003.

[13] A. Oliver, J. Freixenet, R. Martí, J. Pont, E. Pérez, E. R. Denton, and R. Zwigelaar. A novel breast tissue density classification methodology. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):55–65, 2008.

[14] M. Rahman, B. C. Desai, and P. Bhattacharya. Supervised machine learning based medical image annotation and retrieval. In *Image CLEFmed*, pages 692–701, 2005.

[15] J. Suckling. The mammographic image analysis society digital datagram database. *Excerpta Medica International Congress Series*, 1069:375–378, 1994.

[16] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New York, 1995.

[17] D. S. Watkins. *Fundamentals of matrix computations*. John Wileys Sons, 1991.

[18] L. Wei, Y. Yang, R. M. Nishikawa, and M. N. Wernick. Mammogram retrieval by similarity learning from experts. In *IEEE International Conference on Image Processing*, pages 2517–2520. IEEE, October 2006.

[19] J. N. Wolfe. Breast patterns as an index of risk for developing breast cancer. *American Journal of Roentgenology*, 126:1130–1139, 1976.