

MammoSysLesion: a Content-Based Image Retrieval System for Mammographies

Júlia E. E. de Oliveira, Arnaldo de A. Araújo

Department of Computer Science
Federal University of Minas Gerais
Belo Horizonte, MG, Brazil
{julia,arnaldo}@dcc.ufmg.br

Thomas M. Deserno

Department of Medical Informatics
Aachen University of Technology
Aachen, Germany
tdeserno@mi.rwth-aachen.de

Abstract— In this paper, we present a content-based image retrieval system designed to retrieve mammographies from large medical image databases. The system is developed based on breast density and lesion, according to the categories defined by the American College of Radiology, and is integrated to the database of the Image Retrieval in Medical Applications (IRMA) project, that provides images with classification ground truth. Two dimensional principal component analysis is used for texture characterization and support vector machine is used for image retrieval task. Average precision rates are in the range of 72.14% to 80.64% considering a set of 1,392 mammographies.

Medical images database, content-based image retrieval, two-dimensional principal component analysis, breast lesion.

I. INTRODUCTION

Medical images are important for diagnosis purposes as they are related to patient's medical report and pathology. Breast cancer represents one of the main causes of death among women in occidental countries (Brazilian National Cancer Institute)¹, and mammography is the most efficient method for early detection of breast cancer since it shows lesions, like microcalcifications and masses, in their initial phases.

In order to standardize the reports of a mammography, the BI-RADS (Breast Imaging Reporting Data System) atlas was designed by the American College of Radiology². For the information about the decline in sensitivity of mammography with increasing breast density, BI-RADS defines breast density I as almost entirely fatty, density II as scattered fibro glandular tissue, density III as heterogeneously dense tissue and density IV as extremely dense tissue.

Another model of estimation is related to the existence of a lesion and its classification. BI-RADS defines mammography assessment category 0 as need additional imaging evaluation and/or prior mammograms for comparison, category 1 as negative, category 2 as benign finding(s), category 3 as probably benign finding, category 4 as suspicious abnormality (biopsy should be considered), category 5 as highly suggestive of malignancy, and category 6 as known biopsy – proven malignancy.

A visual analysis of mammographies is the basis for radiologists to evaluate and report breast density and the use of computers is increasing in an effort to be used as a second opinion. Content-based image retrieval (CBIR) systems appear as a real possibility to aid radiologists in reducing the variability of their analysis and such a system uses visual information extracted from images to retrieve similar images to one query image. It is important to note that it is not the purpose of a CBIR system to provide diagnosis information of the retrieved images but just present similar images according to a certain pattern. The patient records associated with similar images, however, can be used by the radiologist for computer-assisted diagnosis (CAD) and case-based reasoning [1].

In the context of mammography and breast density, Kinoshita *et al.* used breast density as a pattern to retrieve 1,080 mammographies from the Clinical Hospital from University of São Paulo, Ribeirão Preto, Brazil [2]. 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 79% and 83% were obtained for 50% of recall and precision rates between 79% and 86% were obtained considering the first 25% of the retrieved images.

With the aim of breast lesion classification in mammographies, Verma *et al.* proposed a new algorithm specifically for masses [3]. From DDSM (Digital Database for Screening Mammography) database, regions of interest (ROI) of 100 images of benign cases and 100 images of malignant cases were described using gray level histogram attributes.

It is a challenge for the development of CBIR systems the appropriate characterization of images and the storage and management of the big amount of images produced by hospitals and medical centers. The Image Retrieval in Medical Applications (IRMA)³ 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 [4]. There are currently more than 30,000 diagnostic images with available ground truth information in the IRMA database which are used for CBIR and CAD [5]. Regarding mammography, there are more than 10,000 images in the

¹<http://www.inca.gov.br>

²<http://www.acr.org>

³<http://irma-project.org>

database [6], all of them also with available ground truth information.

In this paper, we propose, implement, and evaluate a CBIR system called MammoSysLesion. A contribution of this work is to introduce the two-dimensional principal component analysis (2DPCA) method [7] for the characterization of breast density texture together with the existence of a breast lesion and its classification, which allows for feature extraction at the same time that dimensionality reduction is performed. Two-dimensional principal component analysis overcomes principal component analysis (PCA) as it is simpler and more straightforward to use for image feature extraction since 2DPCA is directly applied to the image matrix. Retrieval is performed with the aid of a support vector machine (SVM) [8], that is able to solve a variety of learning, classification, and prediction problems.

II. BREAST DENSITY AND LESIONS CHARACTERIZATION

In CBIR systems, the access to information is performed by the visual attributes extracted from images and which can be numerically represented by a feature vector. 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 images analysis. In addition, the process of characterization affects all the subsequent process of a CBIR system [9].

Visually, breasts of different types of tissue and lesion may differ through gray level intensities, and thus their representation can be performed by the texture attribute, since it contains information about the spatial distribution of gray levels and variation in brightness [10]. Nevertheless, the high dimensionality of a feature vector that represents texture may limit its computational efficiency, so it is desirable to choose a technique that combines the representation of texture with the reduction of dimensionality, in a way to turn the retrieval algorithm more effective and computationally treatable. The two-dimensional principal component analysis (2DPCA) technique is able to satisfy these requirements.

2DPCA technique [7] is based on 2D matrices rather than 1D vector like in PCA technique [11], as image covariance matrices can be constructed directly using the original image matrices. The idea of 2DPCA technique is to project image \mathbf{A} , a matrix of size $m \times n$ pixels, onto \mathbf{X} by the linear transformation $\mathbf{Y} = \mathbf{A}\mathbf{X}$. A projected m -dimensional vector \mathbf{Y} is obtained and defined as the projected feature vector of an image \mathbf{A} .

In a way to get a good projection vector \mathbf{X} , the trace of the covariance matrix of the projected feature vectors is obtained through the adoption of the criterion $J(\mathbf{X}) = tr(\mathbf{S}_x)$, where \mathbf{S}_x denotes the covariance matrix of the projected feature vectors of the training examples and $tr(\mathbf{S}_x)$ denotes the trace of \mathbf{S}_x .

Thus, the image covariance matrix \mathbf{G} of an image \mathbf{A} can be defined as $\mathbf{G} = E[(\mathbf{A}-E\mathbf{A})^T(\mathbf{A}-E\mathbf{A})]$. Hence, for a given image \mathbf{A} , let $\mathbf{Y}_k = \mathbf{A}\mathbf{X}_k$, $k = 1, 2, \dots, d$, where d corresponds to the number of selected eigenvalues. A family of projected features $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_k$ is obtained, which is called principal components (vectors) of the

image \mathbf{A} . Unlike PCA, where the principal component is a scalar, with 2DPCA each principal component is vector. The principal component vectors are used to form an $m \times d$ matrix $L = Y_1^T, Y_2^T, \dots, Y_k^T$, which is called the feature matrix or feature image of the image \mathbf{A} .

III. SUPPORT VECTOR MACHINE FOR CONTENT-BASED IMAGE RETRIEVAL

Image retrieval aims at retrieving, from a database, images that are relevant to a given query. The query image goes through the process of feature extraction in order to be compared to the feature vectors of all images stored in the database. The most similar images with respect to the query are retrieved and presented to the radiologist.

Support vector machine (SVM), for a binary classification, can be described as follows: given two classes and a set of points that belong to these classes, the SVM 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 data cannot be precisely separated by a hyperplane, a function called kernel is used instead. It receives two points x_i and x_j from the input space, according to $K(x_i, x_j) = \Phi(x_i) \Phi(x_j)$, and computes the product between these data in the feature space. The most commonly used kernels are the polynomial and the Gaussian.

On the other hand, for more than two classes, the problem turns a multiclass problem [12]. In the one against all method, a SVM is built for each class through the discrimination of this class against the remaining classes. The number of SVMs used in this method is M . A test data b is classified using a decision strategy, i.e., the class with the maximum value of discriminant function $f(b)$ is assigned to that data. All the n training examples are used to construct the SVM for one class. The SVM for one class is built using the set of training data (b) and the desired outputs (y).

In the one against one method, a SVM is built for a pair of classes through its training in the discrimination of two classes. In this way, the number of SVMs used in the method is $M(M-1)/2$. One SVM for a pair of classes is built using training examples belonging to only the two classes.

IV. EXPERIMENTS

The MammoSysLesion system was implemented using Matlab through the image processing toolboxes, and the LIBSVM library [13]. Feature extraction was executed on an IntelCore2Quad 2.66 GHz processor with 8Gb of RAM under Microsoft Windows 64 bits operating system. Image retrieval was performed on an IntelCore2Duo 2 GHz processor with 3Gb of RAM under Microsoft Windows 32 bits.

The mammographies used in this work were selected from the database of radiological images of the IRMA project and were generated using several film digitizers [6]. In the IRMA project, all images are coded according to a mono-hierarchical, multi-axial coding scheme [14], and this codification provides ground truth of all mammographies, as all the images are previously verified by an experienced radiologist. From both cranio-caudal (CC) and medio-lateral (MLO) projections, 1,392 images were selected from IRMA database, and these images are from BI-RADS categories I to IV for breast density and from categories 1, 2, and 5 representing breast lesion. Mammographies from other BI-RADS categories for breast lesion (0, 3, and 4) are present only in a few amount at the database, therefore they were not taken into account for these tests.

In order to select regions of interest (ROI) of the images containing only breast tissue and lesion, excluding artifacts such as annotations and exam labels from mammographies, and as the 2DPCA technique requires that all images have the same size, it was extracted, through an automatic process, ROIs of size 300 x 300 pixels. After this extraction, the methodology applied to the experiments was:

Step 1 → **2DPCA feature extraction**: 2DPCA technique was performed in each of the 1,392 ROIs. The following principal components related to the first d largest eigenvalues of the covariance matrix were used in the experiments: 2, 4, 6, 8, and 10.

Step 2 → **Measurement of similarity between images**: SVM was used to indicate the relevance of images to a certain query. Using the LIBSVM library, the set of 1,392 feature vectors was divided into 60% of the feature vectors for training and 40% of the feature vectors for test. The selection of the feature vectors used for training and the ones used for test was random. Moreover, tests were done using the polynomial and Gaussian kernels.

Step 3 → **Evaluation of the CBIR system**: measures of precision and recall were obtained and the average precision for 10% of recall was chosen, since radiologists pay more attention to the top returned images.

The performance of 2DPCA technique was compared to the ones using principal component analysis (PCA) and singular value decomposition (SVD) for breast and lesions characterization, as these two techniques are also able to represent texture and reduce the dimensionality of the feature vector. SVM was evaluated for the task of image retrieval.

V. RESULTS AND DISCUSSION

Table II lists the average precision, for the selected first d principal components, comparing breast density and lesion characterization using 2DPCA, PCA, and SVD, and with SVM for the retrieval task.

It can be observed that using the polynomial kernel, the 2DPCA technique obtained the highest values of average precision, although very close to the values obtained by the SVD technique, and with the retention of the first 10 principal components, an average precision

of 80.38% was achieved. For the Gaussian kernel, the SVD technique do not overcome the 2DPCA technique only for the first 4 principal components, which obtained an average precision of 80.64%. The texture of the breast density and lesion was best represented by the features extracted using the 2DPCA technique, which in these tests could capture the difference between the gray level intensities of the different breast tissues together with the breast lesion. In spite of this, other tests for the identification of the lesions, especially on dense densities should be carried out. The 2DPCA technique has showed, in general, numerically invariant to the number of principal components, indicating the possibility of reducing the dimensionality of the feature vector and still representing the characteristics of the breast density and lesion through the retention of a lower number of principal components.

For the characterization of the breast density and lesion using the technique 2DPCA and keeping the 4 first principal components, that obtained the best average precision, the time of execution of the CBIR system was 6,200.00 seconds using the polynomial kernel of SVM and 4,3 seconds using the Gaussian kernel. As a CBIR that takes minutes to execute the retrieval process is not viable for the implementation of a system that can be used by radiologists, the use of the polynomial kernel was discarded.

Therefore, using the 2DPCA technique for breast density and lesion characterization and keeping the 4 first principal components, and SVM with the Gaussian kernel, Fig. 1 presents, considering the average precision, the precision and recall curve.

TABLE I. AVERAGE PRECISION FOR THE SELECTED PRINCIPAL COMPONENTS COMPARING 2DPCA, PCA, AND SVD TECHNIQUES FOR BREAST DENSITY AND LESION CHARACTERIZATION AND COMPARING THE POLYNOMIAL AND GAUSSIAN KERNELS OF SVM CLASSIFIER.

d	Technique	Polynomial kernel	Gaussian kernel
2	2DPCA	78.54%	72.14%
	PCA	63.13%	67.92%
	SVD	74.83%	76.03%
4	2DPCA	79.1%	80.64%
	PCA	66.85%	68.54%
	SVD	77.94%	77.44%
6	2DPCA	79.26%	76.14%
	PCA	66.1%	70.23%
	SVD	78.92%	76.24%
8	2DPCA	80.00%	75.34%
	PCA	66.91%	69.72%
	SVD	78.09%	77.92%
10	2DPCA	80.38%	76.88%
	PCA	67.57%	71.17%
	SVD	78.79%	77.49%

With respect to the 2DPCA technique, for 10% of recall, a precision of 83% means that from 56 mammographies retrieved by the system, 47 are relevant for the user query.

Fig. 2 shows an example of the MammoSysLesion system, with a query image of BI-RADS category III for breast density and that have a malignant lesion (BI-RADS category 5) – first image on top and left. With exception of two images, all the retrieved images belong to the same category of breast density of the query image. Although not all the retrieved images have a lesion of the same category of the query image, this happens because malignant lesions appears in mammographies containing more ramifications than benign lesions and both lesions present themselves as brighter regions. A texture attribute was not enough to capture the difference between the lesions and a dense breast density.

VI. CONCLUSION

This work may contribute to the area of content-based image retrieval of mammographies, providing a system able to aid radiologists in their diagnosis or a system that is useful as pre-processing stage for computer-aided systems for breast lesions classification. Experiments were performed in a way to establish the number of principal components necessary to characterize texture at the same time the dimensionality reduction is performed, using the 2DPCA technique.

Future works include the characterization of breast lesions individually, through morphological features. Also, other features can be used to characterize breast density together with the texture attribute.

ACKNOWLEDGMENT

This research 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.

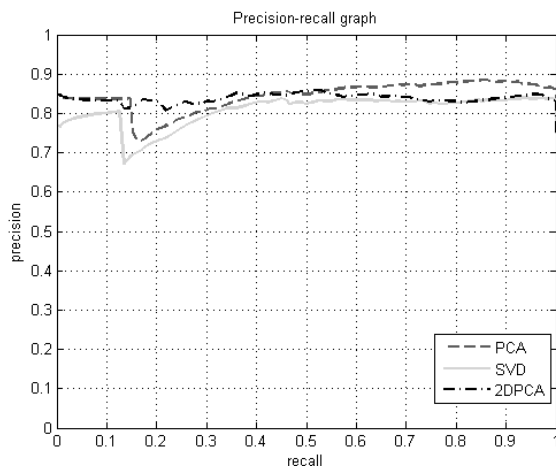


Figure 1. Precision and recall curve comparing 2DPCA, PCA, and SVD techniques for breast density and lesion characterization and SVM with Gaussian kernel for the retrieval process, considering the average precision.

REFERENCES

- [1] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications", *Int. J. Med Inform.*, vol. 73, pp. 1-23, 2004.
- [2] S. K. Kinoshita, P. M. A. Marques, R. R. P. Junior, J. A. H. Rodrigues, and R. M. Rangayyan, "Content-based retrieval of mammograms using visual features related to breast density patterns", *J. Dig. Imag.*, vol. 20, pp. 172-190, 2007.
- [3] B. Verma, "Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammograms", *Artif. Intell. Med.*, vol. 180, pp. 257-262, 2008.
- [4] 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", *Meth. Inform. Med.*, vol. 43, pp. 354-361, 2004.
- [5] 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", *Comput. Med. Imag. Graph.*, vol. 29, pp. 143-155, 2005.
- [6] J. E. E. de Oliveira, M. O. Güld, A. de A. Araújo, B. Ott, and T. Deserno, "Towards a standard reference database for computer-aided mammography", *Proc. of SPIE Med. Imag.*, vol. 6915, pp. 69151Y, 2008.
- [7] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, pp. 131-137, 2004.
- [8] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
- [9] R. Baeza-Yates, and B. R. Neto, *Modern Information Retrieval*. Addison-Wesley Professional, 1999.
- [10] R. C. Gonzales, R. E. Woods, and S. L. Eddings, *Digital Image Processing using Matlab*. Prentice-Hall, 2003.
- [11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley Sons, 2001.
- [12] C. -W. Hsu, and C. -J. Lin, "A comparison of methods for multiclass support vector machines", *IEEE Trans. Neural Netw.*, vol. 13, pp. 415-425, 2002.
- [13] C. -C. Chang, and C. -J. Lin, "LIBSVM: a library for support vector machines, 2001", Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [14] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. Wein, "The IRMA code for unique classification of medical images", *Proc. of SPIE Med.*, vol. 5033, pp. 440-451, 2003.

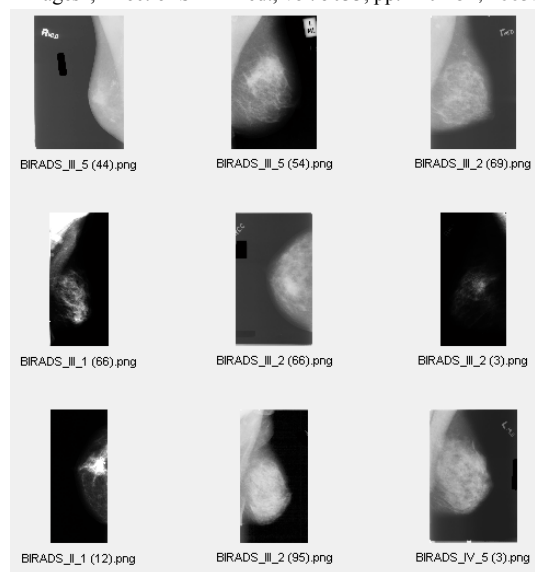


Figure 2. Example of the MammoSysLesion system for image retrieval based on breast density and lesion, using the 4 first principal components of 2DPCA technique and SVM with Gaussian kernel.