

False Positive Reduction in Breast Mass Detection Using Two-Dimensional PCA

Arnau Oliver, Xavier Lladó, Joan Martí, Robert Martí, and Jordi Freixenet

Institute of Informatics and Applications
University of Girona
Campus Montilivi, Ed. P-IV, 17071, Girona, Catalunya, Spain
{aoliver, llado, joanm, marly, jordif}@eia.udg.es

Abstract. In this paper we present a novel method for reducing false positives in breast mass detection. Our approach is based on using the Two-Dimensional Principal Component Analysis (2DPCA) algorithm, recently proposed in the field of face recognition, in order to extract breast mass image features. In mammography, it is well known that the breast density measure is highly related to the risk of breast cancer development. Hence, we also propose to take advantage of a previous breast density classification in order to increase the overall breast mass detection performance. We test our approach using a set of 1792 RoIs manually extracted from the DDSM database. Moreover, we compare our results with several existing methods. The obtained results demonstrate the validity of our approach, not only in terms of improving the performance but being a generalizable, simple, and cost-effective approach.

1 Introduction

Breast cancer is a major health problem in western countries. A study developed in the United States by the American Cancer Society estimates that between one in eight and one in twelve women will develop breast cancer during their lifetime [1]. The most used method to detect breast cancer is mammography, because it allows the detection of the cancer at early stages, a crucial issue for a high survival rate [2].

A breast mass is a localized swelling, protuberance, or lump in the breast, and usually is one of the clearest signs of breast cancer. Recently, several algorithms have been proposed for the automatic detection of masses [3,4]. Most of these algorithms are based on solving two different steps: firstly, the detection of regions with high probability of being mass (Regions of Interest, RoIs); and secondly, a validation step to ensure that the detected RoIs really depict true masses. Note the second step, which is the main interest of this paper, deals with the well-known problem called *false positive reduction*.

Yang et al. [5], in the framework of face detection and face recognition, have recently proposed the Two-Dimensional Principal Component Analysis algorithm (2DPCA) with the goal of improving the eigenfaces method [6]. Both approaches are based on finding (by means of the Karhunen-Loeve transform) the sub-space

which contains the principal variations of the original face image database. Afterwards, the projection of an unknown face sample into the sub-space provides an efficient image representation which is used for face classification purposes. As Yang et al. stated, the use of 2DPCA has important advantages over PCA: firstly, it is simpler and more straightforward to use for image feature extraction since 2DPCA is directly based on the image matrix; secondly, it is easier to evaluate the covariance matrix accurately; and thirdly, it is computationally more efficient.

In this work we present a novel method for reducing false positives in breast mass detection. Our approach is inspired by the 2DPCA algorithm initially proposed for face recognition purposes. While the face recognition framework deals with a large number of different people, including images of different viewpoints and illuminant conditions, in the image mass detection we are dealing with a two-class problem. Note that although several mammographic databases are available with a large number of cases, from the mass detection point of view, they only contain two different types of RoIs: the ones containing masses and the ones containing normal tissue. Therefore, the intra class variability is mainly due to grey-level and texture differences of the breast, and also to the shape and size of the mass or other structures present in the RoIs.

Recent studies have also shown [7] that the sensitivity of most of the mass detection algorithms is decreasing as the density of the breast increases. Therefore, it is difficult to achieve the desired constant sensitivity which is required at a given specificity of these systems. Moreover, the breast density measure is highly related to the risk of breast cancer development [8]. Hence, taking those issues into account, one could argue that the knowledge of the breast tissue prior to the detection could improve the breast mass detection performance. Although in this paper this classification is manually done by an expert, recent works have shown the feasibility of using automatic systems for this purpose [9].

The rest of the paper is structured as follows: Section 2 introduces the 2DPCA algorithm. In Section 3 we describe our proposal of false positive reduction in breast mass detection, while Section 4 describes how to introduce the breast density information into the system. Section 5 shows the obtained results, comparing them with the classical PCA approach and also with different existing methods. Finally, conclusions and further work are discussed in Section 6.

2 The 2DPCA Approach

Both eigenfaces and 2DPCA algorithms were initially designed for face recognition purposes, where an unknown face image is assigned to a known person. These algorithms start with a database of M face images corresponding to I known individuals, where usually $I \ll M$ because the database contains a set of images for each person, including variations on pose and light.

In the original eigenfaces approach each image – of width w and height h – is represented by a 1D vector x_k of length $N = w \times h$ which contains all the grey-level values. Given the database of 1D face images, the Karhunen-Loeve transform is used in order to find the vectors that best account for the distribution of face images within the entire image space (the eigenvectors). However,

the transformation of 2D images to 1D image vectors seems an unnatural trick, as actually the algorithm is dealing with images. This fact was discussed by Yang et al. [5] and originated the main motivation for their 2DPCA proposal. Opposed to conventional PCA, 2DPCA is based on 2D matrices rather than 1D vectors. Therefore, using 2DPCA the image covariance matrix G_t is defined as:

$$G_t = \frac{1}{M} \sum_{j=1}^M (A_j - A_\mu)^t (A_j - A_\mu) \quad (1)$$

where A_μ is the mean image of all training samples. Then, using the Karhunen-Loeve transform it is possible to obtain the corresponding face space, which is the subspace defined as:

$$\begin{cases} \{X_1, \dots, X_d\} = \arg \max |X^t G_t X| \\ X_i^t X_j = 0, \quad i \neq j, \quad i, j = 1, \dots, d \end{cases} \quad (2)$$

where X is a unitary column vector. The first equation looks for the set of d unitary vectors where the total scatter of the projecting samples is maximized (the orthonormal eigenvectors of G_t corresponding to the first d largest eigenvalues). On the other hand, the second equation is needed to ensure orthonormality.

With the selected set of eigenvectors is possible to construct a family of feature vectors for each image. Thus, for an image sample A , the projected feature vectors (the principal components) Y_1, \dots, Y_d are found by:

$$Y_k = AX^k, k = 1 \dots d \quad (3)$$

It is important to note that while for PCA each principal component is a scalar, for 2DPCA each principal component is a vector. It is this set of vectors for image that is used to construct the feature image (a matrix of size $m \times d$) as $B = [Y_1, \dots, Y_d]$.

In a similar way to the eigenfaces approach, comparing images means to compare the constructed features. As the dimension of the feature space has increased in one dimension, now the comparison of images is done by comparing matrices:

$$d(A_1, A_2) = \sum_{k=1}^d \|Y_k^i - Y_k^j\| \quad (4)$$

where $\|Y_k^i - Y_k^j\|$ denotes the Euclidean distance between the two principal components (vectors) Y_k^i and Y_k^j .

3 Mass Detection Using 2DPCA

The transition from face recognition to mass detection is not an easy task, basically due to the variance of the grey-level range of the images and the multiple sizes of the RoIs. Note that this size is depending on the detected mass, and there is a huge range of mass sizes [2].

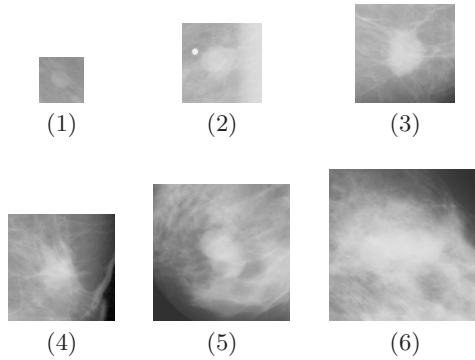


Fig. 1. One ROI with clear mass (the brightest region) belonging to each size group. Note that the relation between the mass lesion and the ROI is maintained.

Grey-level and texture variation of ROIs are mainly related to the variation of the acquisition parameters (exposure time, x-ray energy) of mammograms obtained at different time intervals and also to the nature of the breast (breast density, thickness). Assuming a commonly used simplification, these parameters are considered to affect only the range of the grey-level values of each ROI. Thus, a solution which takes those variations into account is computed by equalizing the images. In this sense, a uniform distribution model is used for equalization.

On the other hand, and in contrast with the face recognition framework where a database of faces of the same size is available, the size of the ROIs is not always the same. In order to deal with this problem of the size, different proposals were analyzed in [10]. The authors concluded that better performances were obtained when the database was clustered in different groups according to their size. Note that in this situation, the mass sizes variability is reduced for each cluster. Thus, when a new ROI has to be analyzed, the corresponding cluster in terms of its size is used. In our work, we also adopt this strategy to solve the problem of the mass size. Figure 1 shows one image sample belonging to each size group (6 different sizes are considered). Each of those groups is then used to apply the 2DPCA algorithm described in Section 2. Thus, the 2DPCA is used to extract breast mass image features according to its size.

4 Including Breast Density Information

As described in Section 1, the sensitivity of most of the mass detection algorithms decreases as the density of the breast increases. In order to take advantage of such information, we introduce a previous step of breast density classification with the goal of increasing the overall breast mass detection. This step consists on a first classification of the database of ROIs according to the breast density parameter. Therefore, we can divide our database of images based on their breast density. Although recent works [9] have demonstrated that this breast density

classification task can be done by automatic procedures, we perform this step manually by means of an expert and following the 4 classes specified in the BIRADS standard [11] widely used in radiology.

In order to evaluate the improvement of this step, we will repeat our experiments of mass detection for both situations: the first one in which the original database of RoIs is directly used; and the second one in which the database of RoIs is previously classified according to the breast density.

5 Experimental Results

The evaluation of our experiments is done using a leave-one-out scheme and Receiver Operating Characteristics (ROC) analysis. In the leave-one-out methodology, we are using for each RoI the set of features extracted by 2DPCA approach. Subsequently, these features are compared with the features obtained with the model, classifying the RoIs according to our two-class problem: RoIs depicting a mass or RoIs depicting normal tissue. This procedure is repeated until all the RoIs have been used as a query image. The classifier used, is a combination of the C4.5 decision tree [12] and the k-Nearest Neighbour algorithm. This classifier provides a numerical value related to the membership of each class. Thus, varying the threshold of this membership it is possible to generate the ROC analysis [13], widely used in the medical field. In such analysis, the graphical curve represents the true positive rate as a function of the false positives rate. Moreover, the percentage value under the curve (known as A_z) is an indication for the overall performance of the observer, and is typically used to analyze the performance of the algorithms.

However, in order to perform a more global evaluation of our results we propose to compute the A_z value for different ratios of number of RoIs depicting masses and number of RoIs depicting normal tissue (from ratio 1/1 to ratio 1/6). The idea of analyzing these different ratios is twofold: firstly, to evaluate the performance of our method on different levels of difficulty; and secondly, to compare our proposal with existing methods (Section 5.3). It is important to notice that previous works only provide results for specific ratios. Hence, analyzing all these ratios will enable the comparison with them. On the other hand, and for showing the improvement of the 2DPCA, we include a previous developed algorithm directly based on the classical eigenfaces approach (from now on, we will refer to it as PCA approach).

The algorithm was evaluated using a database of 1792 RoIs extracted from the DDSM mammographic database [14]. From this set, 256 depicted a true mass, while the rest 1536 were normal, but suspicious tissue. According to the size of the lesion, we used six different groups of RoIs. In order to evaluate in more detail our proposal in terms of using 2DPCA and considering breast density information, we will focus on the experiments with the ratio 1/3 although we include a plot showing the results of all the ratios. Each group of RoIs corresponded to the following mass sizes intervals: $< 10 \text{ mm}^2$, $(10 - 60) \text{ mm}^2$, $(60 - 120) \text{ mm}^2$, $(120 - 190) \text{ mm}^2$, $(190 - 270) \text{ mm}^2$, $> 270 \text{ mm}^2$, and the number of masses in each interval were respectively, 28, 32, 37, 57, 69, and 33 masses.

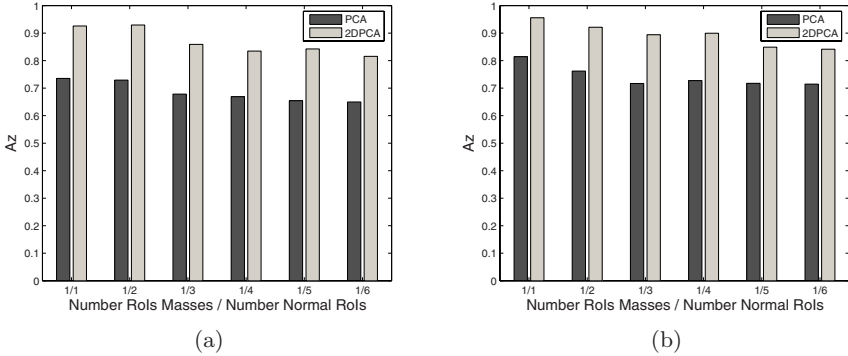


Fig. 2. Performance of our approach using the DDSM database. (a) without including breast density information and (b) considering this information.

5.1 Results Without Considering Breast Density Information

For obtaining the results of this experiment all the RoIs of the database were used, classifying them only according to their size. Figure 2(a) shows the mean Az value obtained using the leave-one-out strategy and varying the ratio between both kind of RoIs. Note the performance of both PCA and 2DPCA approaches decreases as the ratio of RoIs depicting masses decrease. For the PCA approach we obtained $Az = 0.73$ for the ratio 1/1 and $Az = 0.60$ for the ratio 1/6, while using the 2DPCA approach we obtained $Az = 0.92$ and $Az = 0.81$ respectively. Thus, the 2DPCA approach obtained better performances than the PCA.

The Az values for the ratio 1/3 are detailed in the first row of Table 1. The overall performance of the system at this relation is 0.70 for PCA and 0.86 for the 2DPCA. Note that both approaches are more suitable for false positive reduction of larger masses than smaller ones. This is due to the fact that larger masses have a larger variation in grey-level contrast with respect to their surrounding tissue than smaller masses, which are usually more subtle, even for an expert.

5.2 Results Considering Breast Density Information

Figure 2(b) shows the obtained mean Az values considering the breast density information. We classified the RoIs database not only according to the RoIs size, but also to the density of the breast. Using this added knowledge, the performance of both PCA and 2DPCA approaches increased compared with the previous experiment. In particular, for the PCA approach we obtained $Az = 0.81$ for the ratio 1/1 and $Az = 0.71$ for the ratio 1/6, while for the 2DPCA we obtained $Az = 0.96$ and $Az = 0.85$ respectively.

The last two rows of Table 1 show the Az values for the ratio 1/3 according to the size of the mass. Clearly, the overall performance of the approaches increased. For instance, the mean performance of PCA improved up to 0.75 while the 2DPCA performance was 0.91, obtaining an improvement of 5%.

Table 1. Az results (ratio 1/3) for the classification of masses according to the RoI size. Final column shows the mean Az value. S1 to S6 refers to the six RoIs sizes, from small to bigger one.

		Az						Mean
		$S1$	$S2$	$S3$	$S4$	$S5$	$S6$	
Without breast information	PCA	0.53	0.70	0.70	0.68	0.72	0.83	0.70
	2DPCA	0.81	0.83	0.87	0.84	0.89	0.93	0.86
With breast information	PCA	0.70	0.71	0.71	0.72	0.77	0.89	0.75
	2DPCA	0.88	0.93	0.91	0.92	0.89	0.92	0.91

Table 2. Works dealing with mammographic mass false positive reduction, detailing the number of RoIs and the ratio (number of RoIs with masses / number of normal tissue RoIs) used. Further, we include the results obtained with the proposed approach at the same ratio (where BDI means including breast density information).

	Az of Other Works			Az of Presented approaches			
	$RoIs$	$Ratio$	Az	PCA	$2DPCA$	$PCA+BDI$	$2DPCA+BDI$
Sahiner [15]	672	1/3	0.90	0.70	0.86	0.75	0.91
Qian [3]	800	1/3	0.86	0.70	0.86	0.75	0.91
Chang [16]	600	1/1	0.83	0.73	0.92	0.81	0.96
Tourassi [4]	1465	$\cong 1/1$	0.87	0.73	0.92	0.81	0.96

5.3 Discussion

We include in Table 2 a comparison of our method with the performance of various existing methods. Note that our efforts have concentrated on obtaining the same ratio of masses used in their experiments. However, we want to clarify that not all methods used the same databases and therefore our aim is only to provide a general view of the performance of our approach with respect to different strategies. For instance, the works of Sahiner et al. [15] and Qian et al. [3], which used a ratio 1/3, obtained Az values of 0.90 and 0.83 respectively. Note that using the 2DPCA approach with specific density learning, we obtain better performances. Similar behaviour is observed with the proposals which used a ratio of 1/1.

6 Conclusions and Further Work

In this paper, we have presented a new strategy which is a generic, simple and cost-effective method for mass segmentation. The strategy consists on training a classifier with RoIs representing masses and normal tissue, but using different training sets according to the internal breast density category. The classification algorithm is based on the use of 2DPCA for extracting RoIs features.

The performance of the system was evaluated using a leave-one-out methodology and ROC analysis, and calculated at different ratios of RoIs with masses and RoIs depicting normal tissue. The obtained results demonstrate the validity

of our proposal. Further work will be focused in expanding the training database with the aim to detect other kind of mammographic lesions, such as microcalcifications or architectural distortions.

Acknowledgments

This work was partially supported by MEC grant nb. TIN2006-08035.

References

1. American Cancer Society: Breast cancer: facts and figures. 2003-04. ACS (2003)
2. Kopans, D.: Breast Imaging. Lippincott-Raven, Philadelphia (1998)
3. Qian, W., Sun, X., Song, D., Clarke, R.A.: Digital mammography - wavelet transform and kalman-filtering neural network in mass segmentation and detection. *Acad. Radiol.* 8, 1074–1082 (2001)
4. Tourassi, G.D., Vargas-Vorecek, R., Catarious, D.M., Floyd, C.E.: Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information. *Med. Phys.* 30, 2123–2130 (2003)
5. Yang, J., Zhang, D., Frangi, A.F., Yang, J.: Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Machine Intell.* 26, 131–137 (2004)
6. Turk, M.A., Pentland, A.P.: Eigenfaces for recognition. *J. Cogn. Neuro.* vol. 3 (1991)
7. Ho, W.T., Lam, P.W.T.: Clinical performance of computer-assisted detection (CAD) system in detecting carcinoma in breasts of different densities. *Clinical Radiology* 58, 133–136 (2003)
8. Wolfe, J.N.: Risk for breast cancer development determined by mammographic parenchymal pattern. *Cancer* 37, 2486–2492 (1976)
9. Oliver, A., Freixenet, J., Martí, R., Zwiggelaar, R.: A comparison of breast tissue classification techniques. In: Larsen, R., Nielsen, M., Sporning, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 872–879. Springer, Heidelberg (2006)
10. Oliver, A., Martí, J., Martí, R., Bosch, A., Freixenet, J.: A new approach to the classification of mammographic masses and normal breast tissue. *IAPR Int. Conf. on Patt. Rec.* 4, 707–710 (2006)
11. American College of Radiology: Illustrated Breast Imaging Reporting and Data System BIRADS. 3rd edn. American College of Radiology (1998)
12. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, New York (1993)
13. Metz, C.: Evaluation of digital mammography by ROC analysis. In: *Int. Work. on Dig. Mammography*, pp. 61–68 (1996)
14. Heath, M., Bowyer, K., et al.: The Digital Database for Screening Mammography. In: *Int. Work. on Dig. Mammography*, pp. 212–218 (2000)
15. Sahiner, B., Chan, H.P., et al.: Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue. *Med. Phys.* 23, 1671–1684 (1996)
16. Chang, Y.H., Hardesty, L.A., et al.: Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment. *Med. Phys.* 28, 455–461 (2001)