Mass eigendetection and the benefits of introducing breast density information

J. Freixenet, A. Oliver, X. Lladó, R. Martí, J. Martí
Institute of Informatics and Applications, University of Girona, Catalonia (Spain)

Abstract The purpose of this paper is to present a novel algorithm for mass detection in a mammographic computer-aided diagnosis system. Four key points provide the novelty of our approach: (1) the use of eingenanalysis for describing variation in mass shape and size; (2) a Bayesian formulation providing a mathematical sound framework, flexible enough to include additional information; (3) the use of two dimensional PCA for false positive reduction; and (4) the incorporation of breast density information, an internal feature of the breasts closely related to the performance of most mass detection algorithms and which, in contrast, has not been considered in existing approaches. The robustness and the database independence of our approach are shown by the fact that different databases are used for training and testing procedures.

Keywords CAD · Mammography · Mass detection · Breast density

1. Introduction

Breast cancer is one of the most common forms of cancer in women of western countries, remaining the leading cause of death in women aged 40–55 in the United States. Mammography allows the detection of breast cancer at its early stages, a crucial issue for lowering the death rate. Computer aided detection (CAD) systems are being developed to help radiologists to detect and diagnose new cases. However, most of these algorithms do not take the breast density into account, which has been shown as an important factor for the performance of these systems. Usually, as most dense is the breast, worst is the performance of the CAD. In this paper we present a new framework for mass detection based on a template matching scheme which incorporates the breast density information.

2. Methods

Three different steps can be clearly identified in our approach: the template creation, the matching step, and the false positive reduction step. As stated, breast density is also included in all the steps.

2.1 Template creation

The first step of the algorithm is the design of a set of reliable templates. This is based on learning the shapes and sizes from real masses. Thus, the initial input of the algorithm is a database of roughly annotated masses, where only the centre and the maximum radius of the masses are known. Note that, in some cases, this is the only information we can obtain, as there are masses for which is difficult to exactly find their boundary. Therefore, using this set of regions of interest (RoIs), the shapes of the masses are analyzed and learnt by an algorithm inspired on the eigenfaces approach [1].

This algorithm makes use of the Karhunen–Loeve transform to find the set of vectors that best account for the distribution of the initial images. In order to reduce the size variance of our problem, we initially cluster the database in few classes according to the lesion size. Thus, at this point, a set of "eigenmasses" per size are found. Subsequently, the contours of the N eigenmasses containing 95% of variation explanation are extracted and, per each size, an initial image template is created where the brightest pixels represent the pixels with high probability to be a contour of the mass, while the darkest ones represent pixels with low probability (for instance, the central region). This template is finally constructed by adding each contour image weighted by its corresponding eigenvalue.

The interesting point of this template creation is that varying the weight of each eigenmass contour, a new shape can be defined or adapted when looking for a new mass in a mammogram. Thus, we

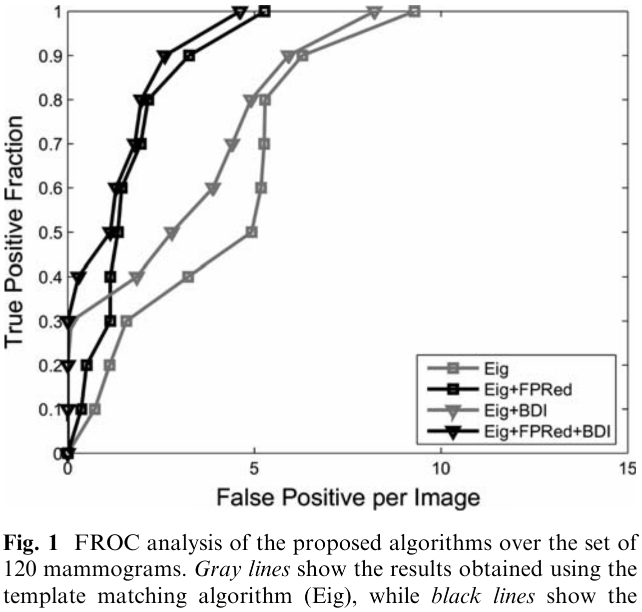

Fig. 1 FROC analysis of the proposed algorithms over the set of 120 mammograms. *Gray lines* show the results obtained using the template matching algorithm (Eig), while *black lines* show the ones obtained by the proposed algorithm and the false positive reduction (Eig + FPRed). *Square marks* are obtained without considering breast density information (*BDI*), and *triangular marks* when this information is included

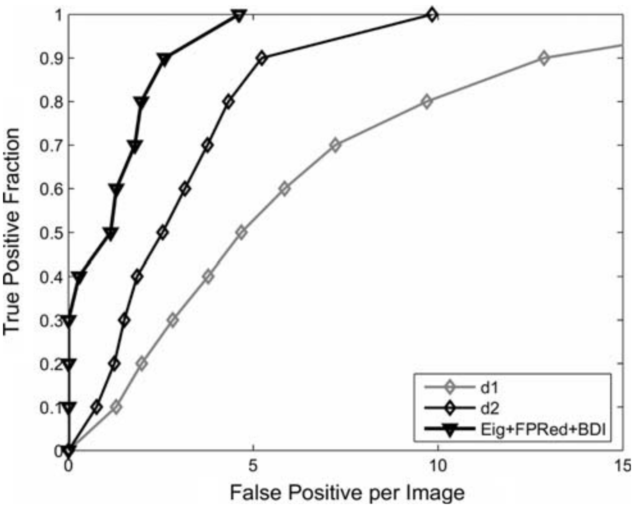

Fig. 2 FROC comparison between the proposed approach and algorithms *d1* and *d2*

will consider as plausible mass shapes, those obtained from normalized linear combinations of the eigenmass contours.

2.2 Matching the template in a mammogram

The second step of the algorithm is the matching process, which is based on a Bayesian pattern matching scheme. We follow the approach of Jain et al. [2], although clear differences are found in the way of specifying the deformations. We define an a priori probability which specifies the global transformations (changes in translation and scale) and local deformations that can be applied to the prototype templates. In contrast to the work of Jain et al. we do not consider rotations for two different reasons: firstly, the created templates are slightly round, and secondly to reduce the computational cost. The likelihood probability is a measurement of the similarity between the templates and the object(s) present in the mammogram. Finally, the a posteriori probability is found by multiplying the a priori and the likelihood probabilities. The aim is to find those points where the a posteriori probability is a maximum. This is approached here by using a gradient ascent algorithm.

The result of this step is a set of regions of interest of the mammogram (RoIs) marked as containing potential masses. However, some of the RoIs actually correspond to normal tissue and therefore, a subsequent step is necessary for reducing the number of false positives.

2.3 2DPCA false positive reduction

The third step is inspired on the 2DPCA algorithm [3], a recent improvement of the eigenfaces algorithm. As stated by the authors, 2DPCA is simpler and more straight-forward to use for image feature extraction since it is directly based on the image matrix, and also it is easier to accurately evaluate the covariance matrix. Note that one could argue that the template in the detection step could be obtained using such algorithm. However, this was not used here for computational reasons. Using our approach, the deformations of the template are modeled using only a vector of coefficients. Instead, with the 2DPCA approach, a vector of vectors will be necessary because, with this approach, each principal component is a vector.

Once the 2DPCA transform is applied, each RoI is compared with the set of original images already transformed (the training set), and it is classified according to the behavior of the most similar image (which can be a mass or normal tissue). Thus, for doing such task, a different training database of RoIs has to be used, as we need RoIs with masses and RoIs being normal tissue (a proportion of one RoI with masses for three RoIs being normal tissue has been selected to both account for the higher variability of normal tissue and for the higher number of normal RoIs found in screening data).

2.4 The breast density role

The breast density information can be naturally incorporated into the previous steps. The training databases are clustered not only in size, but also according to the breast density. In more detail, from the first step we will obtain a set of templates for RoIs belonging to BIRADS I mammograms, another set for RoIs belonging to BIRADS II mammograms, and so on. Therefore, when looking for masses in a new mammogram, firstly this will be automatically classified according to BIRADS standard (see, for example, the work of Oliver et al. [4] for a recent comparison of automatic procedures to perform such task), and subsequently, only the templates of the classified density will be used to find the masses. This process is also used in the false positive reduction step.

3. Experimental results

The robustness of our approach can be stated by the fact that, in contrast with previous works, we used a database for training (the DDSM database [5]) and another one for testing (the MIAS database [6]). The reasons for this are twofold. Firstly, the algorithm needs a huge database of RoIs for learning and training the system, in order to get representative instances of the different shapes and sizes of the masses, which is given by the DDSM database. Secondly, for evaluation purposes we need an accurately segmentation of the lesions, and we have this information from the MIAS database.

Therefore, a set of 120 mammograms were extracted from MIAS database to test the system, 40 showing confirmed masses (the ground-truth accurately marked by an expert) while the remaining 80 were normal mammograms. According to the experts consensus, the breast density of this set was distributed in 35 cases for BIRADS I, 30 for BIRADS II, 30 for BIRADS III, and 25 for BIRADS IV. On the other hand, a total of 1,440 MLO right mammograms were extracted from the DDSM database for training. We selected only MLO mammograms as the MIAS database only shows this view. Among these, 360 present a mass being the rest normal ones.

In Fig. 1 the capability to detect masses of the presented algorithm is evaluated using FROC analysis. The gray line with square marks shows the proposed template matching performance, obtaining a number of false positives per image. This number is clearly reduced by the false positive reduction algorithm, the black line with square marks. The lines with triangular marks are obtained when including the breast density information. Note that the performance of both algorithms is increased when we take the internal breast density into account.

A comparison between our approach and others is provided in Fig. 2. These approaches are based on the works of Tourassi et al. [8] and Karssemeijer [7]. The former (*d1*) is a CBIR based algorithm using mutual information which can be considered as an adaptation of the original work of Lai et al. [9]. The latter (*d2*) is a scale-space based algorithm which extracts a set of features of suspicious regions, and classifies them using a kNN algorithm. Note that our approach clearly outperforms both algorithms.

In terms of ROC analysis, the overall performance over the 40 mammograms containing masses resulted in an area under the curve (Az) of 0.84 ± 0.08 and 0.88 ± 0.06 without and with considering breast density information, respectively. Note that again the benefits of using the breast density information are clear.

4. Conclusions

We have presented a new algorithm for mass detection. We have shown that our proposal, even when using different databases of mammograms for training and testing, outperforms current works. Besides, we have demonstrated the important benefits of developing an initial classification of the mammograms according to their internal tissue. Further work is focusing on the expansion of this approach to a larger database of mammograms in order to clinically assess the full benefits of the developed work.

Acknowledgments

This work was partially supported by MEC grant nb. TIN2006–08035 and by the UdG-IDIBGI grant.

References

1. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
2. Jain A, Zhong Y, Lakshmanan S (1996) Object matching using deformable templates. IEEE Trans Pattern Anal Machine Intel 18(3):267–278
3. Yang J, Zhang D, Frangi, A, Yang J (2004) Two dimensional PCA: a new approach to appearance based face representation and recognition. IEEE Trans Pattern Anal Machine Intel 26(1):131–137
4. Oliver A, Freixenet J, Martí R, Zwiggelaar R (2006) A comparison of breast tissue classification techniques. Proc MICCAI 2:872–879
5. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer P (2000) The digital database for screening mammography. In: Proceedings ofIWDM, pp 212–218
6. Suckling J, Parker J, Dance DR, Astley SM, Hutt I, Boggis CRM et al (1994) The mammographic image analysis society digital mammogram database. In: Proceedings of IWDM, pp 211–221
7. Karssemeijer N, te Brake G (1996) Detection of stellate distortions in mammograms. IEEE Trans Med Imag 15(5):611–619
8. Tourassi G, Vargas-Vorecek R, Catarious D, Floyd C (2003) Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information. Med Phys 30(8):2123–2130
9. Lai SM, Li X, Bischof W (1989) On techniques for detecting circumscribed masses in mammograms. IEEE Trans Med Imag 8(4):377–386