# Mammographic mass eigendetection

Arnau Oliver[a,*], Jordi Freixenet[a], Robert Martí[a], Erika R.E. Denton[b], and Reyer Zwiggelaar[c]

[a]Institute of Informatics and Applications, University of Girona, 17071, Girona, Spain
[b]Department of Breast Imaging, Norfolk and Norwich University Hospital, Norwich, NR4 7UY, UK
[c]Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK

**Abstract.** A new algorithm for the detection of masses in mammographic images is presented. The algorithm has been designed in two steps. Firstly, the regions likely to be a mass are detected by using a deformable template matching approach, where the template is constructed using the eigenimages of a set of manually detected masses. Subsequently, an algorithm adapted from the eigenfaces approach is used to assure that the detected regions really depict true masses (false positive reduction). The evaluation uses a leave-one-out methodology and is based on a database of 120 mammograms, which include 40 masses and 80 normals. ROC and FROC analysis is used to demonstrate the potential of the developed approach.

## 1 Introduction

The proposed method is designed in two steps. Initially, an algorithm is used to detect Regions of Interest (RoI) with a high likelihood of containing a mass and, subsequently, a second algorithm is used for false positive reduction. To our knowledge, this is similar to the strategy used by radiologists: firstly they discard mammograms which certainly do not have a mass present, and in a posterior step, they might re-examine the detected suspicious regions.

Although in the literature different algorithms have been proposed for mass detection [1, 10], few of them incorporate prior knowledge of the shape of the masses. We refer to the work of Lai et al. [4], which looks for masses by using a simple circular template and a pattern matching algorithm using the normalised cross-correlation as the similarity measure. In this work we also use a template matching scheme but with two main differences. Firstly, instead of using the cross-correlation similarity, a probabilistic Bayesian scheme [3] is used. And secondly, contour information coming from an eigenanalysis of manually annotated masses is used, instead of using region information. On the other hand, some techniques have been developed, which are directly related to the false positive reduction step or RoI classification [2, 8]. In contrast with those algorithms which are task-specific, we adapted the eigenfaces approach [9], which is a well known algorithm for automatic face recognition.

## 2 From Eigenfaces to Eigenmasses

The original eigenfaces approach of Turk and Pentland [9] for face recognition is based on the use of the Karhunen-Loeve transform in order to find the vectors that best account for the distribution of face images (forming the face subspace) within the entire image space. The total scatter matrix (the covariance matrix) is calculated as:

$$S_t = \sum_{k=1}^{M} (x_k - \mu)(x_k - \mu)^t \tag{1}$$

where $\mu = \frac{1}{M}\sum_{k=1}^{M} x_k$ is the mean of all face samples and $M$ the number of face images. Using the Karhunen-Loeve transform it is possible to obtain the (dimensionality reduced) eigenfaces space, as the one that maximises:

$$W_{pca} = arg\max_{W} |W^t S_t W| \tag{2}$$

There is significant parallelism between face images of the same person and mammographic RoIs images. As in face images, mammograms are rich-textured images, and it is also possible to talk about variations in illumination and pose. Changes in illumination are related to the different acquisition parameters (dose and energy of X-rays, internal density of the breast), while changes in the pose can be explained as global changes (compression related) or local changes (size and shape of the mass). To prevent these problems we firstly equalise the images using a uniform distribution model, and secondly, we cluster the RoIs of the database according to their size. Note that in contrast with face recognition, the size of the RoIs is not always the same, as it depends on the size of the (possible) mass. Although other proposals can be considered (resizing), we reported in previous work that the used solution outperforms the others [6].

To avoid confusion, note that if the RoI database only contains masses, the result of the application of such algorithm will be called *eigenmasses*, while if the database is composed of RoIs with masses and RoIs with normal tissue, we will call the resulting images *eigenrois*.

*Corresponding author, e-mail: aoliver@eia.udg.es

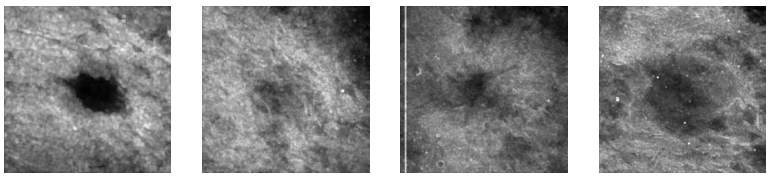

**Figure 1.** The probabilistic templates obtained by clustering the dataset in four clusters of various sizes. Lighter pixels represent a higher probability of a mass contour.

## 3 Probabilistic Mass Contour Template

The first objective of the proposal is the construction of a general mass template, which should take the shape variations into account. The main aim is that pixels with a boundary morphology which has a major representation in the database have a higher probability than pixels not belonging to it. As the template is represented as an image, the pixel brightness is associated with the probability of belonging to a contour. Thus, pixels which certainly do not represent a contour have intensity 0, pixels that always are on a contour (if any) have intensity 1, and the rest of the pixels have intermediate values.

Note that considering only the boundaries of manually segmented masses prefixes a set of contours, and contours different from them will probably be refused while, in contrast, the probability to find two masses with similar shape is very low. Thus, in order to obtain a more general template, it is constructed by looking for the sub-space that these boundaries define. This is achieved adapting the eigenfaces approach described in Section 2. Moreover, using this approach only a rough manual segmentation is needed, just including the centre and size of the mass.

With the obtained eigenmasses, it is possible to construct a probabilistic template per size. For constructing these templates, the $N$ eigenvectors containing $95\%$ of variation explanation were used, considering more probable shapes those with the greatest eigenvalue. Therefore, an initial template is constructed as :

$$\psi^0(x,y) = \frac{1}{N}\sum_{k=1}^{N} w_k W_k(x,y) \tag{3}$$

where $\psi^0(x,y)$ is the template, $W_k(x,y)$ is the $k$-th eigenmass and $w_k$ its normalised eigenvalue (the corresponding eigenvalue divided by their sum). The contour of the eigenmasses is found by extracting the gradient from $\psi^0(x,y)$:

$$\nabla\psi^0(x,y) = \nabla\left\{\frac{1}{N}\sum_{k=1}^{N} w_k W_k(x,y)\right\} = \frac{1}{N}\sum_{k=1}^{N} w_k \nabla W_k(x,y) \tag{4}$$

This equation (image) represents the template as a weighted contours of the eigenmasses. However, in order to discard abnormal shapes, it is necessary to specify the modes of deformation of this template. Plausible shapes are those obtained from linear combinations of the eigenmass contours, and deformation will only affect the weight of the eigenvalues of each eigenmass. This is represented by a vector $\xi$ of size N:

$$\nabla\psi^d(x,y) = \kappa\sum_{k=1}^{N}(\xi_k * w_k)\nabla W_k(x,y) \tag{5}$$

where $\psi^d(x,y)$ is the deformed template and $\kappa$ is just a normalisation factor. With this definition, the vector $\xi$ is all ones when no variations from the template occurs, and results in larger difference to the original template as it increases its values. Hence, assuming a Gaussian distribution, the probability of finding a template with such deformation is:

$$Pr(\xi) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(\xi_k - 1)\} \tag{6}$$

Note that with such definition a new parameter is included. Changes in the value of $\sigma$ represent a more rigid (small $\sigma$) or a more flexible (large $\sigma$) template. Figure 1 shows the templates for four classes representing the range of mass sizes in the database.

## 4 Template Based Detection

Once a template per size range is constructed, the second step is to find mass boundaries (if any) in a mammogram. The developed approach is inspired by the work of Jain et al. [3], where a Bayesian inference scheme was adopted. The three following subsections describe in more detail each of the prior, likelihood and posterior terms.

### 4.1 Prior Distribution

The prior distribution is used to bias the global transformations (changes in translation and scale) and local deformations that can be applied to a prototype template. In contrast with the work of Jain et al. [3], rotation is not taken into account as we assume that this is represented in the probabilistic template. In this work, $\psi^{s,\xi,d}$ denotes a deformation of the original template $\psi^0$. This deformation is performed by locally deforming the template by a set of parameters $\xi$, scaling the local deformation by a factor of $s$, and translating the scaled version along the $x$ and $y$ directions by an amount $d = (dx, dy)$. Assuming that translations and scale sizes have equal probability, and using Eq. 6 for the deformation probability, the prior distribution results in:

$$Pr(s, d, \xi) = K \exp\{-\frac{1}{2\sigma^2} \sum_{k=1}^{N} (\xi_k - 1)\} \tag{7}$$

where $K$ is a normalisation factor. Intuitively, a deformed template with a geometric shape similar to the prototype template is favoured, regardless of its size and location in the image.

### 4.2 Likelihood

The likelihood is a measurement of the similarity between the deformed template and the object(s) present in the image. The deformable template will be attracted and aligned to the salient edges in the input image via a directional edge potential field. For a pixel $(x, y)$ in the input image its edge potential is defined as [3]:

$$\phi_Y(x, y) = -\exp(-\rho\sqrt{\delta_x^2 + \delta_y^2}) \tag{8}$$

where $(\delta_x, \delta_y)$ is the displacement to the nearest edge point in the image, and $\rho$ is a smoothing factor which controls the degree of smoothness of the potential field. This potential is modified by introducing a directional component relating the deformed template $\psi^{s,\xi,d}$ to the edges of the input image $Y$:

$$\Theta(\psi^{s,\xi,d}, Y) = \frac{1}{T} \sum_{x,y \in \psi^{s,\xi,d}} (1 + \phi_Y(x, y)|cos(\beta(x, y))|) \tag{9}$$

where the summation is over all the pixels on the deformed template, $T$ is the number of pixels on the template, $\beta(x, y)$ is the angle between the tangent of the nearest edge and the tangent direction of the template at position $(x, y)$, and the constant 1 is added so that the potentials are positive and take values between 0 and 1. This definition requires that the template boundary agrees with the image edges not only in position, but also in the tangent direction. Using the above energy function, the probability density of the likelihood of observing the input image, given the deformations of the template is:

$$Pr(Y|s, d, \xi) = \alpha \exp\{-\Theta(\psi^{s,\xi,d}, Y)\} \tag{10}$$

where $\alpha$ is a normalising constant to ensure that the above function integrates to 1. The maximum likelihood is achieved when $\Theta(\psi^{s,\xi,d}, Y) = 0$ i.e., when the deformed template $\psi^{s,\xi,d}$ exactly matches the edges in the input image Y.

### 4.3 Posterior Probability Density

Using Bayes rule, the posteriori probability density of the deformed template given the input image is:

$$Pr(s, d, \xi|Y) = Pr(s, d, \xi) * Pr(Y|s, d, \xi)/Pr(Y) \tag{11}$$

where $Pr(Y)$ is the normalisation factor assuring the sum of all probabilities is equal to 1. Using Eqs 7 and 10, the posterior results in:

$$Pr(s, d, \xi|Y) = K_1 \exp\{-\frac{1}{2\sigma^2}[\sum_{k=1}^{N} (\xi_k - 1) + \Theta(\psi^{s,\xi,d}, Y)]\} \tag{12}$$

As the objective is to maximise this probability, we seek to minimise the following objective function with respect to s, $\xi$, d:

$$\Lambda(\psi^{s,\xi,d}, Y) = \sum_{k=1}^{N} (\xi_k - 1) + \Theta(\psi^{s,\xi,d}, Y) \tag{13}$$

This function consists of two terms: a first term that measures the deviation of the deformed template from the prototype, and a second one which describes the fitness of the deformed template to the boundaries of the image.

When the algorithm has finished, a set of points of the mammogram are marked as RoIs. However, a large number of the RoIs actually correspond to normal tissue, and thus, a subsequent step is necessary in order to reduce the number of false positives. Fig. 2 shows the result of applying the algorithm to four well-defined masses.

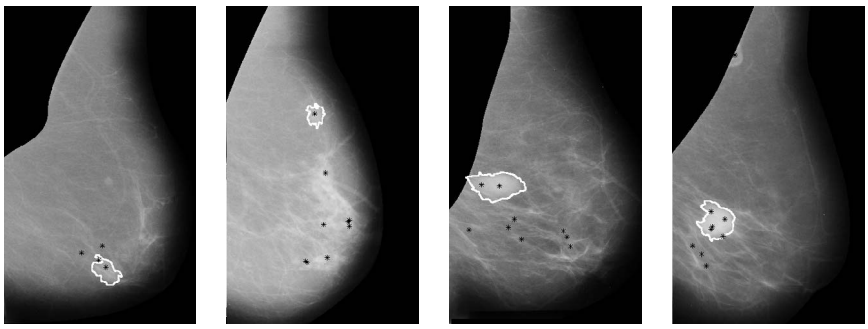

**Figure 2.** Suspicious regions (indicated by the star signs) found in four MIAS mammograms [7], with the annotations of the well-defined masses marked by continuous lines. Note that a posterior step is necessary to reduce the false positive regions.

## 5 False Positive Reduction

Once the most suspicious regions to be a mass have been found, they are cropped using the size of the template, thus obtaining a set of RoIs composed by masses and normal tissue. A classifier is needed to distinguish a RoI according to these two classes, and once again, the eigenfaces approach is used. According to this approach a database of already classified RoIs is necessary. Note that although in this work the RoI database contains only two types of RoIs, this step can be easily extended to distinguish other RoIs containing microcalcifications or architectural distortions. Thus, a model is constructed by using Eq. 2 for each RoI in the database. When a new RoI has to be tested, it will be classified as belonging to the most similar class. Although in the original algorithm the similarity was calculated using the k-Nearest Neighbour algorithm, here we used a combination of this algorithm with the ID3 decision tree. This combination results in a degree of membership for each class (see [6] for more details).

## 6 Results

The performance of our approach was evaluated using a total of $120$ mammograms extracted from the MIAS mammographic database [7]. Of these, $40$ have confirmed masses (containing most of the well-defined and spiculated masses in the database, with detailed manual outline annotations provided by an expert mammographic radiologists) while the rest were normal mammograms. The mammograms containing the well-defined masses are skewed towards the fatty background tissue MIAS classification, whilst the normal and spiculated mass containing mammograms have an even distribution over the background tissue classes. The two sets of RoIs were also extracted from these mammograms. We used four different groups for the $40$ mass RoIs according to their size. Each group corresponds to the following intervals for mass sizes: $< 100\ mm^2, (100 - 180)\ mm^2, (180 - 360)\ mm^2, > 360\ mm^2$. In each interval there were, respectively, 7, 11, 12 and 7 masses, while the rest of RoIs represent normal tissue. Three masses, which were much larger than the rest, were excluded from the modelling. In addition, the original algorithm of Lai et al. [4] (as this is one of the few publications that takes shape information into account when looking for masses) has been re-implemented and included for direct comparison.

The evaluation is done using a leave-one-out methodology and Free Receiver Operating Characteristics (FROC) analysis [5]. In this work, we consider a true positive, if there is at least a $50\%$ overlap between the manually segmented and the automatically detected regions. As is shown in Fig. 3, the proposed approaches have better performances than Lai et al.'s approach, which has a tendency to produce a large number of false positives at high sensitivity rates. Moreover, note that the inclusion of the false positive reduction step clearly reduces the number of false positives detected with the first step of the proposal.

Once the mammograms containing masses are detected, a ROC curve can be obtained measuring the accuracy with which the masses have been detected [5]. The percentage value under the curve ($A_z$) is an indication for the overall performance of the observer, and is typically used to analyse the performance of the algorithms. The overall performance over the $40$ mammograms containing masses resulted in a $A_z$ value of $0.84 \pm 0.11$ and $0.82 \pm 0.13$ without and with the false positive reduction step, while using the Lai et al. algorithm $A_z$ is $0.76 \pm 0.08$. Thus, the proposed approaches have better performance than the original algorithm, although the false positive reduction step introduces a penalisation term in the accuracy of which the algorithm detects the masses. Table $1$ shows the effect of the size of the lesion for the different algorithms. Mean and standard deviation of the values of $A_z$ according to mass sizes are shown. Note that the Lai et al. algorithm has better performances for small masses than for larger ones. In contrast, both proposed approaches have similar performances for each class and significantly better compared to Lai et al.

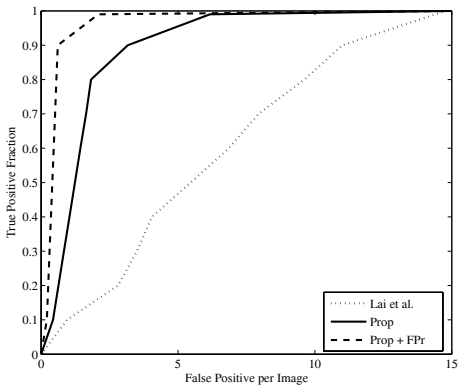


**Figure 3.** FROC analysis of the algorithm over the set of 120 mammograms. It is clear that both proposals outperform the original one.

| | $< 100$ | $100 - 180$ | $180 - 360$ | $> 360$ |
|---|---|---|---|---|
| $A_z$ **(Lai)** | $0.80 \pm 0.05$ | $0.78 \pm 0.08$ | $0.74 \pm 0.09$ | $0.71 \pm 0.05$ |
| $A_z$ **(Eigen)** | $0.85 \pm 0.13$ | $0.83 \pm 0.11$ | $0.82 \pm 0.09$ | $0.87 \pm 0.06$ |
| $A_z$ **(Eigen & FPr)** | $0.83 \pm 0.09$ | $0.81 \pm 0.14$ | $0.80 \pm 0.16$ | $0.80 \pm 0.07$ |

**Table 1.** Influence of the lesion size (in $mm^2$) for the algorithm of Lai et al. and the proposed methods (second row without and third row with false positive reduction (FPr)). The results show mean and standard deviation $A_z$ values.

## 7 Conclusions

The aim of this work is the development of an algorithm capable of finding masses using shape information. The approach is divided into two main algorithms. Firstly, using a database of RoIs containing masses, a probabilistic template is created representing the most probable contours of a mass. This template forms the basis to search for masses in a mammogram using a probabilistic scheme. The result is a set of RoIs containing suspicious regions. Secondly, a false positive classification algorithm has been developed to reduce the number of RoIs not depicting masses. The performance of both approaches was evaluated using a leave-one-out methodology and FROC and ROC analysis. Both stages outperform the work of Lai et al. These results are promising and further work will be focused on: a) a feature selection process to further reduce the number of false positives without penalising the accuracy of the Bayesian template matching, b) the inclusion of information related to the parenchymal tissue distribution (i.e. grey-level and texture information), and c) the use of larger databases to test the clinical relevance of the work.

## Acknowledgments

This work was partially supported by MEC grant nb. TIN2005-08792-C03-01.

## References

1. D. Brzakovic, X.M. Luo, and P. Brzakovic. An approach to automated detection of tumors in mammograms. *IEEE Trans. Med. Imag.*, 9(3):233–241, 1990.
2. Y.H. Chang, L.A. Hardesty, C.M. Hakim, T.S. Chang, B. Zheng, W.F. Good, and D. Gur. Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment. *Med. Phys*, 28(4):455–461, 2001.
3. A.K. Jain, Y. Zhong, and S. Lakshmanan. Object matching using deformable templates. *IEEE Trans. Pattern Anal. Machine Intell.*, 18(3):267–278, 1996.
4. S.-M. Lai, X. Li, and W.F. Bischof. On techniques for detecting circumscribed masses in mammograms. *IEEE Trans. Med. Imag.*, 8(4):377–386, 1989.
5. C.E. Metz. Evaluation of digital mammography by ROC analysis. In *Int. Work. on Dig. Mammography*, pages 61–68, 1996.
6. A. Oliver, R. Martí, J. Martí, A. Bosch, and J. Freixenet. A new approach to the classification of mammographic masses and normal breast tissue. In *IAPR Int. Conf. on Patt. Rec.*, page to appear, Hong Kong, August 2006.
7. J. Suckling, J. Parker, et al. The Mammographic Image Analysis Society digital mammogram database. In *Int. Work. on Dig. Mammography*, pages 211–221, 1994.
8. G.T. Tourassi, R. Vargas-Vorecek, D.M. Catarious, and C.E. Floyd. Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information. *Med. Phys*, 30(8):2123–2130, 2003.
9. M.A. Turk and A.P. Pentland. Eigenfaces for recognition. *J. Cogn. Neuroscience*, 3(1), 1991.
10. R. Zwiggelaar, T.C. Parr, J.E. Schumm, I.W. Hutt, C.J. Taylor, S.M. Astley, and C.R.M. Boggis. Model-based detection of spiculated lesions in mammograms. *Med. Image Anal.*, 3(1):39–62, 1999.