32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Pectoral muscle detection in mammograms based on the shortest path with endpoints learnt by SVMs

Inês Domingues, *Student Member*, *IEEE*, Jaime S. Cardoso, *Member*, *IEEE*, Igor Amaral, Inês Moreira, Pedro Passarinho, João Santa Comba, Ricardo Correia, Maria J. Cardoso

***Abstract*— Automatic pectoral muscle removal on medio-lateral oblique view of mammogram is an essential step for many mammographic processing algorithms. However, the wide variability in the position of the muscle contour, together with the similarity between in muscle and breast tissues makes the detection a difficult task. In this paper, we propose a two step procedure to detect the muscle contour. In a first step, the endpoints of the contour are predicted with a pair of support vector regression models; one model is trained to predict the intersection point of the contour with the top row while the other is designed for the prediction of the endpoint of the contour on the left column. Next, the muscle contour is computed as the shortest path between the two endpoints. A comprehensive comparison with manually-drawn contours reveals the strength of the proposed method.**

I. Introduction

Mammography is the primary imaging modality used for early detection of clinically occult breast cancer. Despite advances in other breast imaging modalities, including ultrasound and magnetic resonance imaging, mammography is still the method of choice.

In medio-lateral oblique (MLO) view mammograms the pectoral muscle is visible as a triangular region of high-density at the upper posterior part of the image. Its presence in mammograms poses an additional source of complexity in automated analysis as it may interfere with the results of image processing methods and induce a bias in breast cancer detection. The texture of the pectoral muscle may also be similar to some abnormalities and may cause false positives in the detection of suspicious masses. Exclusion of the pectoral muscle has thus been taken as an important preprocessing procedure in many mammographic processing methods.

The current image evaluation criteria for the mammographic presentation of the pectoral muscle on the MLO view of the breast recommends that the inferior aspect of the pectoral muscle reaches the level of the nipple. However, many MLO mammograms fail this quality criterion of the image evaluation systems; in [1] it was observed that 75.5% of the mammograms failed the criterion. In some mammograms the pectoral muscle is not present at all. This wide variability in the position of the muscle contour, together with different shapes and intensity contrasts, makes the detection a difficult task.

This work was funded by the Portuguese Innovation Agency (ADI) through project QREN reference 3472 "Semantic PACS". I. Domingues, J. S. Cardoso and I. Amaral are with INESC Porto, Faculdade de Engenharia, Universidade do Porto, Portugal. inesdomingues@gmail.com, jaime.cardoso@inescporto.pt, igor.amaral@gmail.com

I. Moreira, R. Correia and M. J. Cardoso are with Faculdade de Medicina, Universidade do Porto, Portugal. ines.c.moreira@gmail.com, ricardo.jc.correia@gmail.com, mjcard@med.up.pt

P. Passarinho, J. Santa Comba are with Emílio Azevedo Campos, Portugal. {pedro.passarinho, joao.comba}@eacampos.pt

A. Related works

One of the most used pectoral muscle segmentation algorithms is the method proposed by Ferrari et al. [2] based on the Hough transform. The main problem with this approach is that the pectoral muscle is approximated by a line. These methods give poor results when the pectoral muscle contour is a curve. For this reason, the same team proposed another method [3] based on Gabor wavelets. In [4], the pectoral muscle was once again approximated by a straight line, but this line was further adjusted through surface smoothing and edge detection.

Ma et al. [5] described two image segmentation methods: one based on adaptive pyramids and other based on minimum spanning trees. The article [6] chose the longest straight line in Radon-domain as an approximation to the pectoral muscle localization. The problem with this work is two-fold: the simplification of using a line and the use of a private database, so the results cannot be compared with other publications. Camilus and co-workers [7] used a graph cut method followed by Bezier curve smoothing. Recently, an isocontour map methodology was proposed [8]. Finally, a discrete time Markov chain and an active contour model were adopted in [9] for muscle detection.

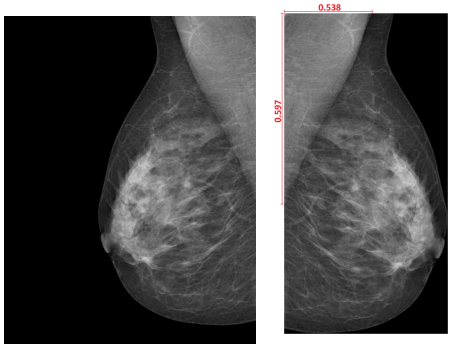
Although the long list of related works, none addresses the problem of deciding if the muscle contour is present or not in the mammogram, all assuming that it is. Moreover, some works assume user input, either in the form of a region of interest or as a set of points in the contour. Finally, with the increased use of digital mammograms, and with its inherent higher quality, simpler approaches could be more adequate.

II. A two step approach for pectoral muscle boundary detection

Techniques based on the shortest path in a graph have been used in the past to address this problem [10]. In particular, interactive approaches based on the livewire method [11] have attracted particular attention. Recognizing the merit of this class of methods, we provide a fully automatic solution. In a first step, the endpoints of the muscle boundary are detected with the help of a regression model; next, the muscle boundary is detected as the shortest path between the now known endpoints.

978-1-4244-4124-2/10/$25.00 ©2010 IEEE

Before pectoral muscle boundary detection, the orientation of the mammograms is automatically judged (by comparing the average intensity level of the right half with the left half) and vertically mirrored those in which the nipple faces to the left so that all pectoral muscles are at the top left corner of the images. Then, an adaptive thresholding method is applied to detect the background; the image is cropped to the bounding box of the breast and muscle regions (see Fig. 1).



(a) Original image. (b) Image after mirroring and cropping.

Fig. 1. Image pre-processing.

A. Automatic detection of the endpoints

The detection of the endpoints is based on techniques of supervised machine learning, namely regression models.

The position of the endpoint on the top row, ET, is normalized by the width of the image (after cropping). Likewise, the position of the endpoint on the left column, EL, is normalized by the height of the image – see Fig. 1(b). Two support vector regression (SVR) models were developed simultaneously. A first SVR model is trained to predict ET, while a second SVR is trained to predict EL.

Support vector machine (SVM) maps data into a high-dimensional space and then classifies them via a hyperplane of maximal margin computed between them [12]. The optimally identified hyperplane in the feature space corresponds to a non-linear decision boundary in the input space. SVR uses the same principles as the SVM for classification, with only a few minor differences such as the replacement of the quadratic error function by an ϵ-insensitive error function, which gives zero error if the absolute difference between the prediction and the target values is less than ϵ, where $\epsilon > 0$. This error function has the key property of producing sparse solutions [12].

The input features chosen to develop the SVR models are the gray-levels values obtained from a 32×32 thumbnail of the cropped mammogram. The SVR model predicting ET is trained with data from the top half of the thumbnail; the SVR model predicting EL is trained with data from the left half of the thumbnail. Therefore the dimension of the feature data is 512 for both models.

B. Pectoral Muscle Boundary Detection with known endpoints

Receiving as input the endpoints predicted by the SVR models, one is left with the computation of the muscle contour between the endpoints. We address this problem by searching the shortest path between the endpoints, after defining a convenient weighted graph in the image.

Intuitively, the muscle boundary manifests itself as a change in the gray-level values of the pixels, giving origin to an edge in the resulting image. Therefore, we can argue that the muscle boundary corresponds to a path through edge pixels. If paths through edges pixels are favored with the appropriate weight in the graph, the muscle boundary is a short path between the two endpoints. Efficient algorithms are available to solve this problem, such as the well-known Dijkstra algorithm [13].

The key steps involved in this operation encompass:

- A gradient computation of the original image. In a broader view, this can be replaced by any feature extraction process that emphasizes the pixels we are seeking for (pixels on the pectoral muscle boundary).
- Consider the gradient image as a weighted graph with pixels as nodes and edges connecting neighboring pixels. Assign to an arc an weight w determined by the gradient values of the two incident pixels.

In this work, the weight of the arc connecting 4-neighbor pixels p and q was expressed as an exponential law:

$$\widehat{f}(g) = f_\ell + (f_h - f_\ell)\ \frac{\exp(\beta\ (255 - g)) - 1}{\exp(\beta\ 255) - 1}, \qquad (1)$$

with $f_\ell, f_h, \beta \in \mathbf{R}$ and g is the minimum of the gradient computed on the two incident pixels. For 8-neighbor pixels the weight was set to $\sqrt{2}$ times that value. The parameters f_ℓ and f_h were fixed at $f_\ell = 2$ and $f_h = 32$; β was experimentally tuned using a grid search method, yielding $\beta = 0.025$.

The gradient model adopted in the experiments reported shortly is based on the Prewitt operator. The Prewitt operator is applied on the x and y directions; from the computed values, G_x and G_y, the magnitude of the gradient is estimated as $z = \sqrt{G_x^2 + G_y^2}$.

C. Proposed Algorithm

The proposed algorithm can be implemented as a sequence of a few high-level operations, as presented in Listing 1. Since the pectoral muscle may be absent in some MLO images, if the endpoints predicted by the SVR models are outside the valid range, the mammogram is assumed without muscle and the shortest path algorithm is not run.

III. RESULTS

The methodology proposed in this paper was assessed on a set of 150 mammograms. One hundred mammograms were collected at Hospital S. João (HSJ), Porto, Portugal. These mammograms are Full Field Digital Mammography (FFDM), already preprocessed by the acquisition equipment

```
Pre-Processing:
   -mirror the image if left half
       is darker than right half
   -crop the image

Main-Processing:
   -compute the endpoints of the contour using
       the previously trained SVR models
   -if either of the endpoints is outside the range [0,1]
       output "muscle is absent" and exit
   -compute the weighted graph
   -compute muscle contour as the
      shortest path between the two endpoints
```

Listing 1: Main operations of the proposed method.

for display purposes. Our detection algorithm was also tested on 50 images from the Digital Database for Screening Mammography (DDSM) (University of South Florida, 2001). In order to evaluate the performance of the proposed method, an experienced radiologist was asked to manually mark the contours on all the 150 mammograms.

Each of the two databases was randomly divided at half in a training set and a test set for the purpose of estimating the prediction accuracy of the SVR model. To select the best parametrization of each model, we conducted a 5-fold cross validation scheme for the training set. The SVR model was based on the SVM package LIBSVM [14], implemented with the epsilon-SVR and a linear kernel. The parameters selected in the HSJ SVR for ET are $\epsilon = 0.01$ and Cost$= 9$ and the parameters for the EL SVR are $\epsilon = 0.06$ and Cost$= 7$. For DDSM data, the ET SVR parameters are $\epsilon = 0.10$ and Cost$= 1$ and the parameters for the EL SVR are $\epsilon = 0.60$ and Cost$= 8$. All the results in this section for the position of the endpoints are obtained with the test set.

To evaluate the quality of the muscle boundary detection algorithm we conducted a complete objective evaluation, based on the hausdorff and the average distances to compare two contours. The hausdorff distance is defined as $H(A,B) = \max(h(A,B), h(B,A))$, where A and B represent the sets of the pixels in the reference muscle contour and the segmented muscle contour respectively, and $h(A,B) = \max_{a \in A} \min_{b \in B} ||a-b||$ and $||.||$ is the Euclidean distance.

Fig. 2 shows the evolution of the absolute error when estimating the endpoints position of the muscle contour. A trend (not shown) was observed for predicted ET to be on the left, and predicted EL to be above Ground Truth points. Table I summarizes the results. Next, we evaluated the

TABLE I

OVERALL RESULTS FOR OVER THE TWO DATABASES, IN THE POSITION OF THE ENDPOINTS.

Database	endpoint top row	endpoint left column
DDSM	0.0598	0.1012
HSJ	0.0943	0.1193

quality of the shortest path to find the muscle contour with known endpoints. Fig. 3 shows the evolution of the error when estimating the position of the muscle contour. The error in pixels was normalized by the size of the diagonal of the

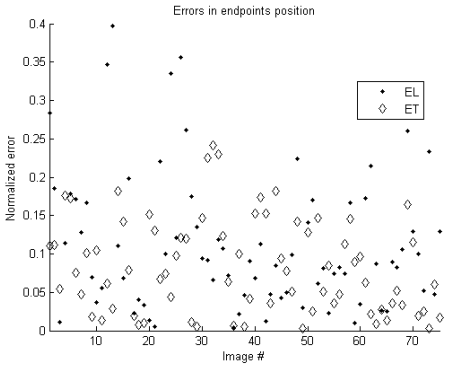


Fig. 2. Error in the position of the endpoints of the proposed method over 50 HSJ mammograms and 25 DDSM mammograms.

image. Table II summarizes the results. To understand if the

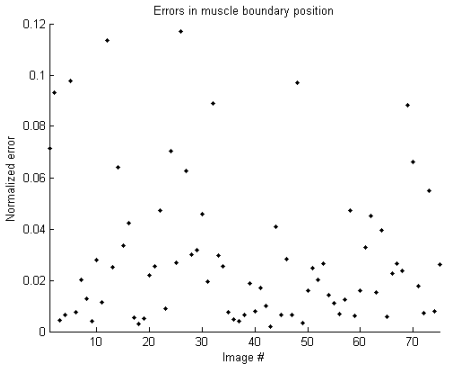


Fig. 3. Error (normalized by the diagonal of the image) in the position of the muscle boundary of the proposed method as measured by the average distance over 50 HSJ mammograms and 25 DDSM mammograms.

TABLE II

OVERALL RESULTS FOR OVER THE TWO DATABASES, IN THE POSITION OF THE MUSCLE BOUNDARY.

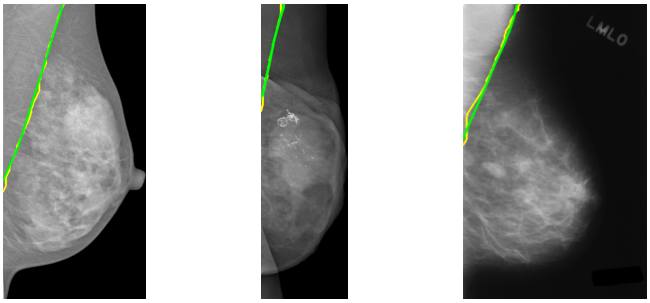
Database	Hausdorff Distance (normalized)	Mean Distance (normalized)
DDSM	0.0860	0.0266
HSJ	0.1232	0.0340

main source of errors was from the endpoint prediction or from the shortest path method, we re-run the shortest path method using as input the true endpoints obtained from the reference contours; results are reported in Table III. Fig. 4 shows some of the images in which the algorithm worked satisfactorily. Some of the unsuccessful cases, displayed in Fig. 5, bring to light the limitations of the current state of the proposed approach. The results in Table III and Fig. 5 seem to indicate that if a robust estimation of the endpoints can be achieved, then the pectoral muscle boundary can be effectively predicted using the shortest path. In fact, from the results, the prediction of the endpoints seems to be the main source of errors for the global estimation of the contour.

TABLE III

OVERALL RESULTS FOR OVER THE TWO DATABASES, IN THE POSITION OF THE MUSCLE BOUNDARY, USING AS ENDPOINTS THE TRUE VALUES FROM THE MANUAL CONTOURS.

Database	Hausdorff Distance (normalized)	Mean Distance (normalized)
DDSM	0.0249	0.0099
HSJ	0.0242	0.0048



(a) Image HSJ #18. (b) Image HSJ #43. (c) Image DDSM #65.

Fig. 4. Selected successful results. Ground truth contours in green and prosed method contours in yellow.

A comparison with literature is not straightforward due to the use of different databases and metrics. Moreover, although not using the same database, in [6] an average hausdorff distance of 12 mm was obtained. After units conversion, our method has an average hausdorff distance of 7 mm.

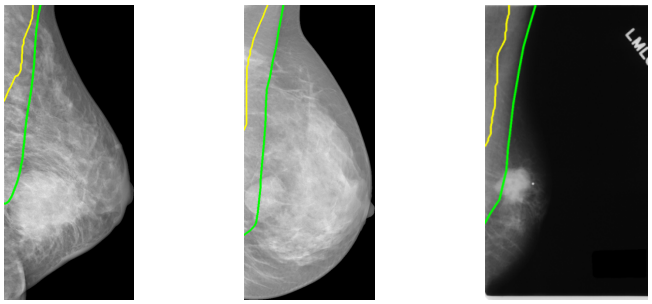
A little surprisingly, the results are not significantly superior over the digital mammograms than over the screen film mammograms. The small size of the databases precludes stronger conclusions.

IV. CONCLUSION

We have presented a method for identification of pectoral muscle contour based on the estimation of the endpoints with a supervised learning algorithm, followed with the delineation of the contour using a shortest path technique. The model learning the endpoints was based on a SVR model, fed with information from a thumbnail of the mammogram.

Initial investigations performed support further work in this direction, as the method achieves a good accuracy. An initial observation is that, since the SVR is using only a thumbnail of the mammogram, the accuracy of the predictions will always be limited by the amount of scaling. Instead of assuming that the shortest path has to start and end on the detected endpoints, we plan to use a window centered on the endpoints or a probabilistic approach to guide the position of the endpoints of the shortest path.

Another, more simple solution, is to learn the shortest path between two full margins of the image, after a convenient coordinate transformation [15]. Although the accuracy obtained with this simpler approach is lower, it has the advantage of simplicity and the potential of generalizing better for equipments from different manufacturers. Further investigation is necessary to assess the trade-offs involved.



(a) Image HSJ #12. (b) Image HSJ #26. (c) Image DDSM #69.

Fig. 5. Selected poor results. Ground truth contours in green and prosed method contours in yellow.

REFERENCES

[1] K. Bentley, A. Poulos, and M. Rickard, "Mammography image quality: analysis of evaluation criteria using pectoral muscle presentation," *Radiography*, vol. 14, no. 3, pp. 189–194, 2008.

[2] R. J. Ferrari, R. M. Rangayyan, J. E. Desautels, R. A. Borges, and A. F. Frère, "Segmentation of mammograms: identification of the skin-air boundary, pectoral muscle, and fibro-glandular disc," in *5th International Workshop on Digital Mammography*, 2000, pp. 573–579.

[3] ——, "Automatic identification of the pectoral muscle in mammograms," *IEEE Trans on Medical Imaging*, vol. 23, no. 2, pp. 232–245, 2004.

[4] S. M. Kwok, R. Chandrasekhar, and Y. Attikiouzel, "Automatic pectoral muscle segmentation on mammograms by straight line estimation and cliff detection," in *IEEE Trans on Medical Imaging*, 2004, pp. 1129–1140.

[5] F. Ma, M. Bajger, J. P. Slavotinek, and M. J. Bottema, "Two graph theory based methods for identifying the pectoral muscle in mammograms," *Pattern Recognition*, vol. 40, no. 9, pp. 2592–2602, 2007.

[6] S. K. Kinoshita, P. Azevedo-Marques, R. R. Pereira, J. A. Rodrigues, and R. M. Rangayyan, "Radon-domain detection of the nipple and the pectoral muscle in mammograms," *Journal Digital Imaging*, vol. 21, no. 1, pp. 37–49, 2008.

[7] K. S. Camilus, V. K. Govindan, and P. S. Sathidevi, "Computer-aided identification of the pectoral muscle in digitized mammograms," *Journal Digital Imaging*, 2009.

[8] B. W. Hong and B. S. Sohn, "Segmentation of regions of interest in mammograms in a topographic approach," *IEEE Trans on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 129–139, 2010.

[9] L. Wang, M. Zhu, L. P. Deng, and X. Yuan, "Automatic pectoral muscle boundary detection in mammograms based on markov chain and active contour model," *Journal of Zhejiang University - Science C*, vol. 11, no. 2, pp. 111–118, 2010.

[10] A. Rebelo, A. Capela, J. P. da Costa, C. Guedes, E. Carrapatoso, and J. S. Cardoso, "A shortest path approach for staff line detection," in *Third International Conference on Automated Production of Cross Media Content for Multi-channel Distribution*, 2007, p. 7985.

[11] A. Chodorowski, U. Mattsson, M. Langille, and G. Hamarneh, "Color lesion boundary detection using live wire," *Proceedings of SPIE*, vol. 5747, pp. 1589–1596, 2005.

[12] V. N. Vapnik, *Statistical Learning Theory*. New York: John Wiley & Sons, 1998.

[13] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[14] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15] J. S. Cardoso, I. Domingues, I. Amaral, I. Moreira, P. Passarinho, J. S. Comba, R. Correia, and M. J. Cardoso, "Pectoral muscle detection in mammograms based on polar coordinates and the shortest path (accepted)," in *32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010.