MASS SEGMENTATION IN MAMMOGRAMS: A CROSS-SENSOR COMPARISON OF DEEP AND TAILORED FEATURES

Jaime S. Cardoso[1]*, *Nuno Marques*[1]

[1]INESC TEC and University of Porto
Porto
Portugal

Neeraj Dhungel[2], *G. Carneiro*[3]†, *A. P. Bradley*[4]

[2]ECE, The University of British Columbia, Canada
[3]ACVT, The University of Adelaide, Australia
[4]ITEE, The University of Queensland, Australia

ABSTRACT

Through the years, several CAD systems have been developed to help radiologists in the hard task of detecting signs of cancer in mammograms. In these CAD systems, mass segmentation plays a central role in the decision process. In the literature, mass segmentation has been typically evaluated in a intra-sensor scenario, where the methodology is designed and evaluated in similar data. However, in practice, acquisition systems and PACS from multiple vendors abound and current works fails to take into account the differences in mammogram data in the performance evaluation.

In this work it is argued that a comprehensive assessment of the mass segmentation methods requires the design and evaluation in datasets with different properties. To provide a more realistic evaluation, this work proposes: a) improvements to a state of the art method based on tailored features and a graph model; b) a head-to-head comparison of the improved model with recently proposed methodologies based in deep learning and structured prediction on four reference databases, performing a cross-sensor evaluation. The results obtained support the assertion that the evaluation methods from the literature are optimistically biased when evaluated on data gathered from exactly the same sensor and/or acquisition protocol.

Index Terms— Mammogram, mass segmentation, transfer learning, cross-sensor

1. INTRODUCTION

The most common imaging modality in breast cancer screening is mammography. Computer-aided detection (CAD) systems have been developed in order to provide a "second reading" to aid the radiologist in reaching a final assessment [1]. A fundamental stage in typical CAD systems is the segmentation of masses in regions of interest (ROIs), which could have been manually or automatically detected. Depending on the application, it may be necessary to have a very precise extraction of the contour of the mass. In the mass classification case, for instance, it is advantageous to have a good segmentation step since some characteristics are extracted from the mass shape (e.g., it is known that convex masses tend to be more benign than very spiculated masses).

Most of the state-of-the-art methods rely on level set methods [2, 3] that are usually based on shape and appearance priors that rarely capture all the variation found in breast masses given the strong assumptions made by such models (e.g., strong edges, similar grey value, etc.). Other recently proposed methods are graph-based models with inference procedures that search for optimal paths in the graph. Within these, the closed path approach [4] compared favorably with methods based on active contours, convergence filters and a graph-cut method with a star shape prior. More recently, the task of mass segmentation has been addressed with deep learning methods [5], in some cases cascading them for improved results [6].

However, all the works present several limitations, particularly in terms of the empirical estimation of the performance. First, the lack of reliable datasets and benchmarks makes the evaluation difficult and the comparison of the different methodologies unreliable. Second, the methods tend to be trained and tested for the same dataset, making it unclear how they will perform is general. This is exacerbated by the fact that the few available datasets include mammograms already processed for presentation (Presentation Intent Type field in the DICOM header; these images are intended for viewing by an observer) in the Picture and Archiving Communication System (PACS). This processing is PACS-specific and therefore methods developed under a specific processing may not work in mammograms that have undergone a different processing. Third, both the deep learning based methods and the traditional tailored based features present limitations and advantages that are important to understand for cross-fertilization of ideas.

*This work was funded by the Project "NanoSTIMA: Macro–to–Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE–01–0145–FEDER–000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

†G. Carneiro acknowledges the support by the Australian Research Council Discovery Project DP 140102794. We also thank Nvidia for the TitanX provided for this research project.

The major contributions of this paper are: a) improvements to the tailored, graph-based methodology in [4], by preprocessing the mammogram with a total variation approach and improving the computation of the cost function inputted to the closed path extraction; b) the comparative performance analysis in a cross-sensor scenario of the following state of the art models: the proposed improved version of [4], the conditional random field and structured support vector machines methodologies proposed by Dhungel et al. [6].

2. STATE-OF-THE-ART MODELS FOR MASS SEGMENTATION

We selected two reference models from the literature, both presenting state of the art performance, but adopting quite different technical solutions.

2.1. Closed Path Approach

The closed contour computation is usually addressed by transforming the image into polar coordinates, where the closed contour is transformed into an open contour between two opposite margins, but [4] solves the problem in the original coordinate space. After defining a directed acyclic graph appropriate for this task, the authors address the main difficulty in operating in the original coordinate space, which is that small paths collapsing in the seed point are naturally favored. This issue is addressed [4] by modulating the cost of the edges to counterbalance this bias. In the mass segmentation task, the weights in the graph are set as

$$w = f_l + (f_h - f_l)\frac{\exp((255-d)\beta)-1}{\exp(255\beta)-1}, \qquad (1)$$

where d is the magnitude of the radial derivative in the pixel (normalized in the range [0, 255]), and f_l, f_h and β are set to 2, 32 and 0.025, respectively. Eq. (1) defines a exponentially monotonous decreasing function between the derivative and the cost, with f_l being the lowest cost and f_h the highest cost assigned in the graph.

2.2. Deep Learning and Structured Prediction Based Approaches

The deep learning and structured prediction based approaches proposed by Dhungel et al. [6] consist of two probabilistic graphical models, namely 1) Structured support vector machines (SSVM) and 2) Conditional Random Field (CRF). Both of these models includes a number of deep learning based shape models, such as patch based deep belief networks (DBN) [7], convolutional neural network (CNN) [8] based on global image along with other models based on Gaussian mixture model (GMM) [9] and shape prior. The SSVM model uses graph cuts [10] for inference and cutting plane optimization [11] to learn the parameter of the model. Similarly, the inference of the CRF model is based on Tree re-weighted belief propagation (TRW) [12, 13] and the parameters of the CRF model are learned with truncated fitting algorithm [12].

3. IMPROVED CLOSED PATH APPROACH

In this section, we propose several improvements to the closed path approach presented in Sec. 2.1, as illustrated in Fig. 1. In particular, we introduce a preprocessing step based on total variation to denoise and enhance the ROI data.

Given an input 2D signal $\mathbf{x}$, the goal of total variation regularization [14] is to find an approximation, $\mathbf{y}$, that is "close" to $\mathbf{x}$ but has smaller total variation than $\mathbf{x}$. One measure of closeness is the sum of square errors and the total variation of $\mathbf{y}$ can be defined by

$$V(\mathbf{y}) = \sum_{n,m} \{|\mathbf{y}_{n+1,m} - \mathbf{y}_{n,m}| + |\mathbf{y}_{n,m+1} - \mathbf{y}_{n,m}|\}$$

So the total variation denoising problem amounts to minimize the following discrete functional over the signal $\mathbf{y}$:

$$0.5\sum_{n,m}(\mathbf{x}_{n,m} - \mathbf{y}_{n,m})^2 + \lambda V(\mathbf{y}) \qquad (2)$$

Furthermore, while the reference methodology [4] relied only in the radial derivative $g(p)$ at a pixel p to set the weights in the graph, we consider both the radial derivative and a measure of regularity of the grey values, both inside and outside the mass. For each pixel p, we consider the radial segment through p connecting the centre $\mathcal{O}$ to the border of the ROI. We compute the standard deviation of the gray values, $std_i(p)$ and $std_e(p)$, in the segments connecting $\mathcal{O}$ to p and p to the border pixel, respectively. The function $h(p) = \max(std_i(p), std_e(p))$ measures the quality of the pixel p in terms of internal and external regularity. Finally, the fitness of p is measured by

$$f(p) = g(p)^a / h(p)^b, \qquad (3)$$

where a and b are to be set experimentally. The weight in the graph is finally set as in the original work [4], see Eq.(1), but with β also set experimentally.

4. CROSS-SENSOR EXPERIMENTAL ANALYSIS

We performed the standard intra-sensor analysis, by splitting the data from a single database in two parts, one for training and the other for performance estimation. Additionally, we also study the cross-sensor performance by training and testing in different databases. We use Dice metric, D, to assess the segmentation accuracy:

$$D = 2\frac{\#(X \cap Y)}{\#X + \#Y},$$

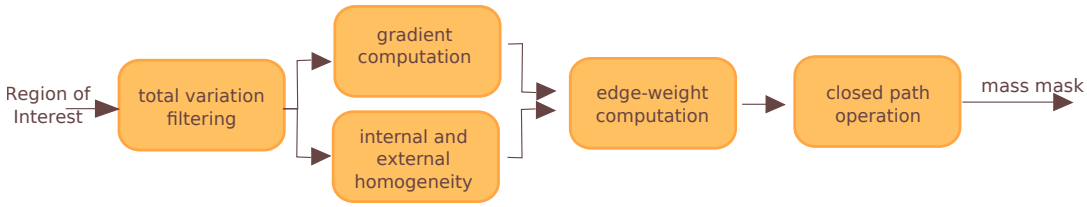


Fig. 1: Block diagram of the improved closed path approach.

where X is the annotated mass region, $\#X$ is the number of pixels in X, Y is the detected mass region, $\#Y$ is the number of pixels in Y.

4.1. The Databases

We conduct our experimental work in four databases, INBreast [15], DDSM-BCRP [16], and two subsets from BCDR [17]. The INBreast comprises Full Field Digital Mammographies (FFDM) acquired in Porto, Portugal, between April 2008 and July 2010; the acquisition equipment was the MammoNovation Siemens FFDM, with a solid-state detector of amorphous selenium, pixel size of 70 μm (microns), and 14-bit contrast resolution. The image matrix was 3328×4084 or 2560×3328 pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). Images were saved in the DICOM format. The INBreast database provides 116 masses with high-quality manual annotation by experts with the OsiriX software.

The DDSM-BCRP [16] database consists of 39 cases (77 annotated images) for training and 40 cases (81 annotated images) for testing, with rough manual annotation of the masses. These film mammograms were scanned on a HOWTEK 960 digitizer with a sample rate of 43.5 microns at 12 bits per pixel.

The BCDR database is organized in four subsets, two with film based mammograms and two with digital mammographies. We selected BCDR-F02 (film based) and BCDR-D01 (digital) for the transfer learning evaluation. BCDR-F02 includes 188 masses manually annotated while BCDR-D01 comprises 143 masses. For the film based subset, MLO and CC images are grey-level digitized mammograms with a resolution of 720 (width) by 1168 (height) pixels and a bit depth of 8 bits per pixel, saved in the TIFF format. In the digital subset, the MLO and CC images are grey-level mammograms with a resolution of 3328 (width) by 4084 (height) or 2560 (width) by 3328 (height) pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). The bit depth is 14 bits per pixel and the images are saved in the TIFF format.

The rectangular Region of Interests (ROI) for our experiments were generated from the bounding boxes (BB) of annotated mass, by expanding the BB by 20%. For the closed path approach, the seed point was set at the centre of the ROI.

Table 1: Mass segmentation on Mammograms: Intra-sensor results. Results are the mean of the Dice metric (the higher the better).

Database	Original Closed Path	Improved Closed Path	SSVM	CRF
INBreast	0.88	0.89	0.90	0.90
BCDR-D01	0.84	0.87	0.88	0.89
BCDR-F02	0.72	0.77	0.83	0.82
DDSM-BCRP	0.52	0.87	0.90	0.90

Table 2: Mass segmentation on Mammograms: Cross-sensor results. Results are the mean of the Dice metric (in brackets is the decrease from the intra-sensor performance).

Train Database	Test Database	Improved Closed Path	SSVM	CRF
BCDR-D01	INBreast	0.89 (0.00)	0.82 (0.08)	0.81 (0.09)
BCDR-F02	INBreast	0.83 (0.06)	0.88 (0.02)	0.87 (0.03)
DDSM-BCRP	INBreast	0.83 (0.06)	0.87 (0.03)	0.87 (0.03)
INBreast	BCDR-D01	0.87 (0.00)	0.82 (0.06)	0.81 (0.08)
BCDR-F02	BCDR-D01	0.84 (0.03)	0.80 (0.08)	0.79 (0.10)
DDSM-BCRP	BCDR-D01	0.84 (0.03)	0.84 (0.04)	0.83 (0.05)
INBreast	BCDR-F02	0.75 (0.02)	0.77 (0.06)	0.80 (0.02)
BCDR-D01	BCDR-F02	0.75 (0.02)	0.77 (0.06)	0.76 (0.06)
DDSM-BCRP	BCDR-F02	0.77 (0.00)	0.81 (0.02)	0.81 (0.01)
INBreast	DDSM-BCRP	0.65 (0.22)	0.77 (0.12)	0.81 (0.09)
BCDR-D01	DDSM-BCRP	0.65 (0.22)	0.83 (0.07)	0.81 (0.09)
BCDR-F02	DDSM-BCRP	0.87 (0.00)	0.85 (0.05)	0.83 (0.07)

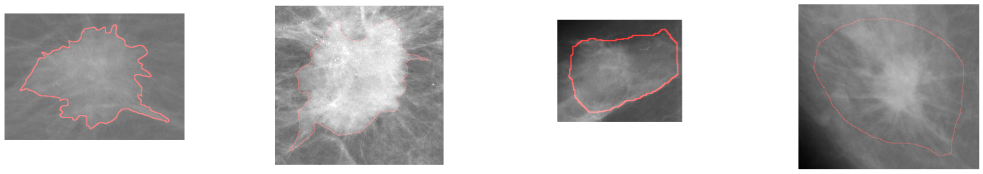
For the intra-sensor setting evaluation, half of the masses in a database were selected for training and the other half for testing. In the cross-sensor experiment, all masses in the source database were used in training, and all masses in the target database were used for testing.

4.2. Results

Results are presented in Table 1 and Table 2. A first clear conclusion is the improvement of the closed path method. While the original method [4] performs well in INBreast, its performance decreases significantly in the other databases, specially in DDSM and BCDR-F02. The improved version has a much more robust behaviour, showing a drop in performance only in the BCDR-F02 database.

The worst performances are obtained when transferring from INBreast to DDSM and from BCDR-D01 to BCDR-F02. One of the reasons behind this performance drop lies in the annotation differences between those databases.

The limitations of the DDSM database in terms of the



(a) INBreast. (b) BCDR_D01. (c) BCDR_F02. (d) DDSM.

Fig. 2: Examples of masses and corresponding manual annotations.

manual annotations of the masses are well-known. Instead of following the boundary of the mass, the annotation is often just a blob, typically completely containing the mass, but also a lot of non-mass tissue (see Fig. 2d). The training in this database sets more weight to the shape prior in the models, reinforcing the low relation of the manual boundary with the true boundary of the mass. A similar behavior is observed in the BCDR subsets, but less pronounced. The annotations in the BCDR database are much more accurate than in DDSM, as shown in Fig. 2b and Fig. 2c. Additionally, the results improve from the film based to the digital mammography, which suggests that the higher data quality of the digital mammograms pays off in the segmentation task. Finally, the fine-detailed segmentation of the (digital) INBreast database yields the best automatic segmentation model.

A comparison of the automatic segmentation methods allows us to conclude that the deep learning based methods present better performance in roughly two thirds of the experimental evaluations, showing superior performance. Nevertheless, the performance loss in the cross-sensor scenario was smaller for the closed path method.

5. DISCUSSION AND CONCLUSIONS

This paper discusses and compares three methods for mass segmentation in mammograms, for the yet unexplored cross-sensor setting. The first model uses tailored features and computes the boundary as the optimal closed path in a graph model. The second and third models are based on deep learning features, combined with CRF and SSVM for parameter estimation. The results shown a good performance in general, specifically in the intrasensor scenario. Although the performance remained appreciable, in some cross sensor cases the performance loss was more than 10%.

There are two clear differences between the selected databases: full field digital vs. digitized film (which is a resolution and dynamic range issue) and accurate vs. blobby segmentations (which is a contour length/complexity issue). If we accept that the manual segmentations in both BCDR databases are of similar quality, the performance loss when transferring between both BCDR datasets is essentially due to the differences between the film and digital data. When transferring to the DDSM is results are due not only to the differences in the data but also to the annotations. The higher losses in the latter scenario suggest that the resolution/dynamic range issue is "manageable" once we come to a consensus on how clinicians should annotate mass lesions.

Do we need more training data or better models? For mass segmentation, the answer seems to be yes to both questions. We need more datasets, better annotated datasets (some of the annotations are enough to develop detection algorithms but not segmentation methods) and better models. Based on our analysis, we conjecture that the greatest gains in mass segmentation performance will happen when these needs are addressed.

6. REFERENCES

[1] Silvia Bessa, Ines Domingues, Jaime S. Cardoso, Pedro Passarinho, Pedro Cardoso, Vitor Rodrigues, and Fernando Lage, "Normal breast identification in screening mammography: a study on 18 000 images," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014.

[2] J. Ball and L. Bruce, "A digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation," in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 4973–4978.

[3] P. Rahmati, A. Adler, and G. Hamarneh, "Mammography segmentation with maximum likelihood active contours," *Medical Image Analysis*, vol. 16, pp. 1167–1186, 2012.

[4] Jaime S. Cardoso, Ines Domingues, and Helder P. Oliveira, "Closed shortest path in the original coordinates with an application to breast cancer," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, 2015.

[5] Neeraj Dhungel, Gustavo Carneiro, and Andrew P. Bradley, "Deep structured learning for mass segmentation from mammograms," *CoRR*, vol. abs/1410.7454, 2014.

[6] Neeraj Dhungel, Gustavo Carneiro, and Andrew P. Bradley, *Deep Learning and Structured Prediction for the Segmentation of Mass in Mammograms*, pp. 605–612, Springer International Publishing, 2015.

[7] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks.," in *NIPS*, 2012, vol. 1, p. 4.

[9] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[10] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *TPAMI*, vol. 23, no. 11, pp. 1222–1239, 2001.

[11] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun, "Large margin methods for structured and interdependent output variables," in *JMLR*, 2005, pp. 1453–1484.

[12] Justin Domke, "Learning graphical model parameters with approximate marginal inference," *arXiv preprint arXiv:1301.3193*, 2013.

[13] Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky, "Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching," in *Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics Np, 2003, vol. 21, p. 97.

[14] Antonin Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imaging Vis.*, vol. 20, no. 1-2, pp. 89–97, Jan. 2004.

[15] Ines Moreira, Igor Amaral, Ines Domingues, Antonio Cardoso, Maria J. Cardoso, and Jaime S. Cardoso, "Inbreast: Towards a full field digital mammographic database," *Academic Radiology*, vol. 19, pp. 236–248, 2012.

[16] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer Jr, "The digital database for screening mammography," in *Proceedings of the 5th international workshop on digital mammography*, 2000, pp. 212–218.

[17] Daniel C. Moura and Miguel A. Guevara López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, no. 4, pp. 561–574, 2013.