TREE RE-WEIGHTED BELIEF PROPAGATION USING DEEP LEARNING POTENTIALS FOR MASS SEGMENTATION FROM MAMMOGRAMS

Neeraj Dhungel[†] *Gustavo Carneiro*[†] *Andrew P. Bradley*[⋆] [*]

[†] ACVT, School of Computer Science, The University of Adelaide
[⋆] School of ITEE, The University of Queensland

ABSTRACT

In this paper, we propose a new method for the segmentation of breast masses from mammograms using a conditional random field (CRF) model that combines several types of potential functions, including one that classifies image regions using deep learning. The inference method used in this model is the tree re-weighted (TRW) belief propagation, which allows a learning mechanism that directly minimizes the mass segmentation error and an inference approach that produces an optimal result under the approximations of the TRW formulation. We show that the use of these inference and learning mechanisms and the deep learning potential functions provides gains in terms of accuracy and efficiency in comparison with the current state of the art using the publicly available datasets INbreast and DDSM-BCRP.

***Index Terms*—** Mammograms, mass segmentation, tree re-weighted belief propagation, Deep learning , Gaussian Mixture model.

1. INTRODUCTION

Breast cancer is the most frequent cancer among women (25% of all diagnosed cancers) and the second most common cancer in the world population [1]. Screening mammograms (see Fig. 1) is one of the most effective tools in the early diagnosis of breast cancer, where clinicians look for suspicious masses (among other structures, such as micro-calcifications) [2]. Usually, these mammograms are manually analysed, even though computer aided diagnostic (CAD) systems have shown potential to improve the trade off between sensitivity and specificity commonly observed in this manual analysis [3]. We believe that one of the issues preventing the realization of this potential is the fact that the most of the current state-of-the-art approaches rely on active contours methods [2, 4] that produce sub-optimal results because of their non-convex cost functions and reliance on contour and appearance priors, which cannot represent well the shape and appearance variations observed in the data.

Our proposed methodology explores statistical learning methods using a conditional random field (CRF) model for the segmentation of breast masses from mammograms. The novelties of our approach is the use of recently developed inference procedure called Tree re-weighted belief propagation (TRW) with the supervised learned features produced by deep learning mechanisms. The inference procedure based on TRW is derived from a variational formulation to find the marginals of the CRF model, and the learning process minimizes directly the segmentation error by back propagating the model parameters [5]. The main reason behind the use of TRW is the fact that it has been found to outperform other inference mechanisms, such as graph cuts [6], for low-connectivity graphs (such as the 4-connected graph, which is the connectivity used in our paper) [7]. We also propose the use of several potential functions in this CRF model based on deep learning methods [8], which are able to directly extract image features from mammograms (i.e.,the features are automatically learned instead of being hand-designed). The primary motivation of using deep learning is the fact that it is presenting state of the art results in recent challenging object detection and segmentation problems in computer vision [9], and we believe these result results can be extended to medical image analysis. Given that these statistical models learn all parameters from manually annotated data and that we do not make any assumptions about the shape and appearance of masses, we believe that our proposed approach is capable of modeling all shape and appearance variations encountered in the data if enough annotated training data is available. We test our methodology on the publicly available datasets INbreast [10] and DDSM-BCRP [11], and our methodology produces competitive results in terms of accuracy and with respect to efficiency, our approach is significantly faster than any of the published methods.

[*]This work was partially supported by the Australian Research Council's Discovery Projects funding scheme (project DP140102794). Prof. Bradley is the recipient of an Australian Research Council Future Fellowship(FT110100623)

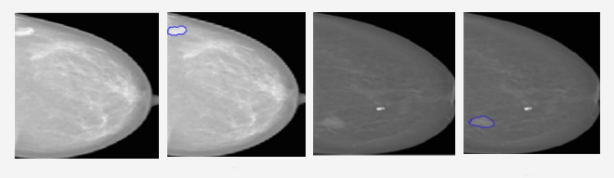

Fig. 1. Examples of mass segmentation from mammograms.

2. METHODOLOGY

In this section, we first describe our model, then the learning and inference methods, which are followed by an explanation of the potential functions.

2.1. Statistical Model for Breast Mass Segmentation

Let $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^N$ be a collection of mammograms, with $\mathbf{x} : \Omega \rightarrow \mathbb{R}$ (Ω denotes the image lattice) representing the region of interest (ROI) containing the mass, and $\mathcal{Y} = \{\mathbf{y}_n\}_{n=1}^N$ representing the mass segmentation of $\mathbf{x}_n$, with $\mathbf{y} : \Omega \rightarrow \{-1, +1\}$ (where $+1$ represents mass and -1, background). We are interested in modeling the probability of a mass annotation $\mathbf{y}$ given an image $\mathbf{x}$, which is represented by a undirected graph with $\mathcal{V}$ nodes and $\mathcal{E}$ edges between nodes, defined as follows [5, 12, 13]:

$$P(\mathbf{y}|\mathbf{x}; w) = \exp\{E(\mathbf{y}, \mathbf{x}; \mathbf{w}) - A(\mathbf{x}; \mathbf{w})\} \quad (1)$$

where $\mathbf{w}$ represents the model parameters,

$$E(\mathbf{y},\mathbf{x};\mathbf{w}) = \sum_{k=1}^{K}\sum_{i\in\mathcal{V}} w_{1,k}\phi^{(1,k)}(\mathbf{y}(i),\mathbf{x}) + \sum_{l=1}^{L}\sum_{i,j\in\mathcal{E}} w_{2,l}\phi^{(2,l)}(\mathbf{y}(i),\mathbf{y}(j),\mathbf{x}) \tag{2}$$

with $\phi^{(1,k)}(.,.)$ representing one of the K potential functions between label (hidden) and pixel (observed) nodes, $\phi^{(2,l)}(.,.,.)$ denoting one of the L potential functions on the edges between label nodes, $\mathbf{w} = [w_{1,1},...,w_{1,K},w_{2,1},...,w_{2,L}]^\top \in \mathbb{R}^{K+L}$ and $\mathbf{y}(i)$ is the i^{th} component of vector $\mathbf{y}$, and $A(\mathbf{x};\mathbf{w}) = \log\sum_{\mathbf{y}\in\{-1,+1\}^{M\times M}}\exp\{E(\mathbf{y},\mathbf{x};\mathbf{w})\}$ is the log-partition function that ensures normalization.

2.2. Tree Re-weighted Belief Propagation

The main issue involving the learning of the model parameters $\mathbf{w}$ in (1) is the computation of the log-partition function $A(\mathbf{x};\mathbf{w})$. A relatively recent approach to solving this problem is based on the use of the variational problem [13] that provides an upper bound to this log-partition function, leading to the design of tree-reweighted belief propagation algorithms that can solve this optimization problem. In particular, the log partition function can be represented as [13]:

$$A(\mathbf{x};\mathbf{w}) = \max_{\mu\in\mathcal{M}} \mathbf{w}^T\mu + H(\mu), \tag{3}$$

where $\mathcal{M} = \{\mu' : \exists\mathbf{w},\mu' = \mu\}$ denotes the marginal polytope, $\mu = \sum_{\mathbf{y}\in\{-1,+1\}^{M\times M}} P(\mathbf{y}|\mathbf{x};\mathbf{w})f(\mathbf{y})$ (with $f(\mathbf{y})$ denoting the set of indicator functions of possible configurations of each clique and variable in the graph [14]), and $H(\mu) = -\sum_{\mathbf{y}}\mathbf{P}(\mathbf{y}|\mathbf{x};\mathbf{w})\log\mathbf{P}(\mathbf{y}|\mathbf{x};\mathbf{w})$ is the entropy. For general graphs with cycles, the marginal polytope $\mathcal{M}$ is difficult to charaterize and the entropy $\mathbf{H}(\mu)$ is not tractable [5]. Tree re-weighted belief propagation (TRW) solves these issues by first replacing the marginal polytope with a superset $\mathcal{L}\supset\mathcal{M}$ that only ensures local constraints of the marginals, and then approximating the entropy calculation with an upper bound. Specifically,

$$\mathcal{L} = \{\mu : \sum_{\mathbf{y}(c)\setminus\mathbf{y}(i)}\mu(\mathbf{y}(c)) = \mu(\mathbf{y}(i)), \sum_{\mathbf{y}(i)}\mu(\mathbf{y}(i)) = 1\} \tag{4}$$

replaces $\mathcal{M}$ in (3) and represents the local polytope (with $\mu(\mathbf{y}(i)) = \sum_{\mathbf{y}'}P(\mathbf{y}'|\mathbf{x},\mathbf{w})\delta(\mathbf{y}'(i)-\mathbf{y}(i))$ and $\delta(.)$ denotes the Dirac delta function), c indexes a graph clique, and the entropy approximation (that replaces $H(\mu)$ in (3)) is defined as

$$\widetilde{H}(\mu) = \sum_{\mathbf{y}(i)} H(\mu(\mathbf{y}(i))) - \sum_{\mathbf{y}(c)}\rho_c I(\mu(\mathbf{y}(c)), \tag{5}$$

where $H(\mu(\mathbf{y}(i))) = -\sum_{y(i)}\mu(\mathbf{y}(i))\log\mu(\mathbf{y}(i))$ is the univariate entropy of variable $\mathbf{y}(i)$, $I(\mu(\mathbf{y}(c))) = \sum_{\mathbf{y}(c)}\mu(\mathbf{y}(c))\log\frac{\mu(\mathbf{y}(c))}{\prod_{i\in c}\mu(\mathbf{y}(i))}$ is the mutual information of the cliques in our model, and ρ_c is a free parameter that if selected properly gives the true upper bound on the entropy.

The estimation of $A(\mathbf{x};\mathbf{w})$ and associated marginals in (3) are achieved via a message passing algorithm, with the following message-passing updates [5]:

$$m_c(\mathbf{y}(i)) \propto \sum_{\mathbf{y}(c)\setminus\mathbf{y}(i)} \exp(\frac{1}{\rho_c}\phi_c(\mathbf{y}(i),\mathbf{y}(j);\mathbf{w}) \prod_{j\in c\setminus i}\exp(\frac{1}{\rho_c}\phi_i(\mathbf{y}(i),\mathbf{x};\mathbf{w}))\frac{\prod_{d:j\in d} m_d(y(j))^{\rho_d}}{m_c(y(j))}, \tag{6}$$

where $\phi_i(\mathbf{y}(i),\mathbf{x};\mathbf{w}) = \sum_{k=1}^{K} w_{1,k}\phi^{(1,k)}(\mathbf{y}(i),\mathbf{x})$ and $\phi_c(\mathbf{y}(i),\mathbf{y}(j);\mathbf{w}) = \sum_{l=1}^{L} w_{2,l}\phi^{(2,l)}(\mathbf{y}(i),\mathbf{y}(j),\mathbf{x})$. Once the message passing algorithm has converged [13], the beliefs for the associated marginals are written as:

$$\mu_c(\mathbf{y}(c)) \propto \frac{1}{\rho_c}\phi_c(\mathbf{y}(i),\mathbf{y}(j))\prod_{i\in c}\phi_i(\mathbf{y}(i),\mathbf{x};\mathbf{w})\frac{\prod_{d:j\in d} m_d(\mathbf{y}(j))^{\rho_d}}{m_c(\mathbf{y}(i))}$$
$$\mu_i(\mathbf{y}_i) \propto \exp(\phi_i(\mathbf{y}(i),\mathbf{x};\mathbf{w}))\prod_{d:i\in d} m_d(\mathbf{y}(i))^{\rho_d} \tag{7}$$

Finally, in order to learn $\mathbf{w}$ in (1), we follow the learning methodology developed by Domke [5], which is based on the minimization of certain loss functions. In particular, we minimize the smoothed univariate classification error, defined as follows:

$$L(\mathbf{w},\mathbf{y}) = \sum_{\mathbf{y}(i)} S(\max_{\hat{\mathbf{y}}(i)\neq\mathbf{y}(i)}\mu(\mathbf{y}(i);\mathbf{w}) - \mu(\hat{\mathbf{y}}(i);\mathbf{w})) \tag{8}$$

where $S(t) = (1+\exp(-\alpha t))^{-1}$ and α controls the approximation quality. This learning process is achieved with truncated fitting of the weight parameter $\mathbf{w}$ with inference using backpropagation in TRW [5].

2.3. Potential Functions

It is worth noticing that the model in (1) can incorporate a large number of different types of potential functions. In this paper, the potential functions between label and pixel nodes are based on deep belief networks (DBN), Gaussian mixture model (GMM), and the mean shape from the training images. The DBN potential function is defined as [8]:

$$\phi^{(1,1)}(\mathbf{y}(i),\mathbf{x}) = -\log P_d(\mathbf{y}(i)=1|\mathbf{x}_S(i),\theta_{d,S}), \tag{9}$$

where $\mathbf{x}_S(i)$ is a patch extracted around image lattice position i of size $S\times S$ pixels, $\theta_{d,S}$ represents the DBN parameters (below, we drop the dependence on $\theta_{d,S}$ for notation simplicity). The DBN model consisting of a network containing Q layers is denoted by:

$$P(\mathbf{x}_S(i),\mathbf{y}(i),\mathbf{h}_1,...,\mathbf{h}_Q) = P(\mathbf{h}_Q,\mathbf{h}_{Q-1},\mathbf{y}(i))\left(\prod_{q=1}^{Q-2}P(\mathbf{h}_{q+1}|\mathbf{h}_q)\right)P(\mathbf{h}_1|\mathbf{x}_S(i)), \tag{10}$$

where $\mathbf{h}_q\in\mathbb{R}^{|q|}$ represents the hidden variables at layer q containing $|q|$ nodes. The first term in (10) is defined by:

$$-\log(P(\mathbf{h}_Q,\mathbf{h}_{Q-1},\mathbf{y}(i))) \propto -\mathbf{b}_Q^\top\mathbf{h}_Q - \mathbf{a}_{Q-1}^\top\mathbf{h}_{Q-1} - \mathbf{a}_y^\top[\tfrac{\mathbf{y}(i)+1}{2},\tfrac{1-\mathbf{y}(i)}{2}]^\top - \mathbf{h}_Q^\top\mathbf{W}\mathbf{h}_{Q-1} - \mathbf{h}_Q^\top\mathbf{W}_y[\tfrac{\mathbf{y}(i)+1}{2},\tfrac{1-\mathbf{y}(i)}{2}]^\top \tag{11}$$

where $\mathbf{a},\ \mathbf{b},\mathbf{W}$ are the network parameters, and the conditional probabilities are factorized as $P(\mathbf{h}_{q+1}|\mathbf{h}_q) = \prod_{i=1}^{|q+1|}P(\mathbf{h}_{q+1}(i)|\mathbf{h}_q)$ because the nodes in layer $q+1$ are independent from each other given $\mathbf{h}_q$, which is a consequence of the DBN structure ($P(\mathbf{h}_1|\mathbf{x}_S(i))$ is similarly defined). Furthermore, each node is activated by a sigmoid activation function $\sigma(.)$, which means that $P(\mathbf{h}_{q+1}(i)|\mathbf{h}_q) = \sigma(\mathbf{b}_{q+1}(i) + \mathbf{W}_i\mathbf{h}_q)$. As a result, (9) is computed by:

$$P_d(\mathbf{y}(i)=1|\mathbf{x}_S(i)) \propto P_d(\mathbf{y}(i)=1,\mathbf{x}_S(i)) = \sum_{\mathbf{h}_1}...\sum_{\mathbf{h}_Q}P_d(\mathbf{x}_S(i),\mathbf{y}(i)=1,\mathbf{h}_1,...,\mathbf{h}_Q), \tag{12}$$

which is estimated with the mean field approximation of the values in layers $\mathbf{h}_1$ to $\mathbf{h}_{Q-1}$ followed by the computation of free energy on the top layer [8]. The learning of the DBN parmeters $\theta_{d,S}$ in (9) is achieved with an iterative layer by layer training of auto-encoders using contrastive divergence [8]. The GMM potential function is defined by:

$$\phi^{(1,2)}(\mathbf{y}(i),\mathbf{x}) = -\log P_g(\mathbf{y}(i)=1|\mathbf{x}(i),\theta_g), \quad (13)$$

where $P_g(\mathbf{y}(i)=1|\mathbf{x}(i),\theta_g) = (1/Z)\sum_{m=1}^{G}\pi_m\mathcal{N}(\mathbf{x}(i);\mathbf{y}(i)=1,\mu_m,\sigma_m)P(\mathbf{y}(i)=1)$ with $\theta_g = [\pi_m,\mu_m,\sigma_m]_{m=1}^{G}$, $\mathcal{N}(.)$ is the Gaussian function, Z is the normalizer, $\mathbf{x}(i)$ represents the pixel value at image lattice position i, and $P(\mathbf{y}(i)=1)=0.5$. The parameter vector θ_g in (13) is automatically learned with the expectation-maximization (EM) algorithm [15] using the annotated training set. The mean shape is computed from the average annotation (estimated from the training set) at each image lattice position $i \in \Omega$, as follows:

$$\phi^{(1,3)}(\mathbf{y}(i),\mathbf{x}) = -\log P_p(\mathbf{y}(i)=1|\theta_p), \quad (14)$$

where $P(\mathbf{y}(i)=1|\theta_p) = \lambda(1/N)\sum_n \delta(\mathbf{y}_n(i)-1)+(1-\lambda)$, where $\lambda \in [0,1]$.

The potential functions between label nodes in (1) encode label and contrast dependent labelling homogeneity. In particular, the label homogeneity is defined by:

$$\phi^{(2,1)}(\mathbf{y}(i),\mathbf{y}(j),\mathbf{x}) = 1-\delta(\mathbf{y}(i)-\mathbf{y}(j)), \quad (15)$$

with $\delta(.)$ denoting the Dirac delta function, and the contrast dependent labelling homogeneity is as follows:

$$\begin{aligned}\phi^{(2,1+n)}(\mathbf{y}(i),\mathbf{y}(j),\mathbf{x}) =& (1-\delta(\mathbf{y}(i)-\mathbf{y}(j))) \\ & \delta(\min(0,\|\mathbf{x}(i)-\mathbf{x}(j)\|_2-\tau_n))\end{aligned} \quad (16)$$

where $\mathbf{x}(i)$ represents the pixel value at position i, and $\tau_n \in \{\tau_1,\tau_2,...,\tau_{10}\}$ is a set with 10 thresholds [5]. Therefore, in total there are 11 pairwise potentials.

3. EXPERIMENTS

In this section, we first present the material and methods used in the experiments, and then we show a comparison between the results produced by our methodology and by other approaches.

3.1. Materials and Methods

The evaluation of our methodology is performed on two publicly available datasets: Inbreast [10] and DDSM-BCRP [11].The Inbreast [10] dataset comprises a set of 56 cases containing 116 accurately annotated masses. We divide this dataset into mutually exclusive train and test sets,containing 58 images each on training and testing sets. The DDSM-BCRP [11] dataset consists of 9 cases (77 annotated images) for training and 40 cases (81 annotated images) for testing. It is worth mentioning that the annotations provided with DDSM-BCRP are generally inaccurate [16, 10], so most of the literature uses subsets of DDSM with bespoke annotations that are not publicly available. Note that some cases in DDSM-BCRP and INbreast database contain multiple masses, and each case presents the Craniocaudal (CC) and Mediolateral (MLO) views.

Segmentation accuracy is assessed with Dice index (DI) = $\frac{2TP}{FP+FN+2TP}$, where TP denotes the number of mass pixels correctly segmented, FP the background pixels falsely segmented as mass, and FN the mass pixels not identified. Efficiency is estimated with the running time of the segmentation algorithm, reported as the average execution time per image on a standard computer (Intel(R) Core(TM) i5-2500k 3.30GHz CPU with 8GB RAM). The ROI to be segmented is obtained by extracting a rectangular bounding box from around the center of the manual annotation from the test/train image, where the size for each dimension of the rectangle is produced by the size of the annotation plus two pixels [17]. This ROI is then resized to 40 x 40 pixels using bicubic interpolation. We use the preprocessing method by Ball and Bruce [4] in order to increase the contrast of the input image.

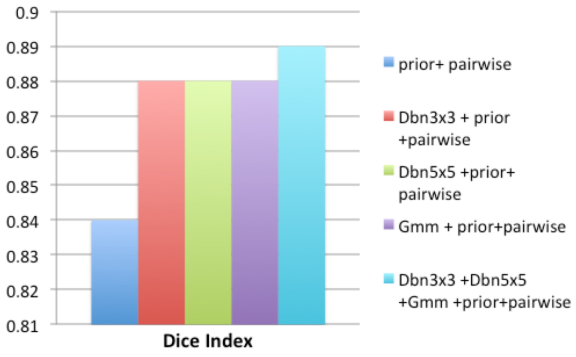

Fig. 2. Dice index over the test set of Inbreast of different versions of our model, containing various subsets of the potential functions.

3.2. Results

The first experiment presented in Fig. 2 shows the importance of the potential functions in the model (1), where we train the model with the Inbreast train set and show the mean Dice index results on its test set. In particular, we show these results using several subsets of the potential functions presented in Sec. 2.3. In this figure, "DBN" and "GMM" represent the potentials $\phi^{(1,k)}$ for $k=\{1,2\}$ with 3×3 and 5×5 denoting the image patch size used by the DBN, and pairwise potentials. It is important to mention that the Dice index of our methodology using all potential functions on the training set is 0.90, which is similar to the performance on the test set shown in Fig. 2. Furthermore, the Dice index of our methodology on the test set when we do not adopt the pre-processing described by Ball and Bruce [4] is 0.85

Tab. 1 shows the Dice index and running time results of our approach (using the model (1) with potential functions DBN3x3 + DBN5x5 + GMM + pairwise) on the test sets of DDSM-BCRP and INbreast. The results from the other methods are as reported by Horsh et al.[16] or by their original authors. The great majority of papers published in this area have used subsets of the DDSM dataset and manual annotations that are not publicly available. For this reason, a direct comparison with these methods is impossible, and we indicate the reproducibility of each experiment in the column 'rep' in Tab. 1. Finally, Fig. 3 shows a few segmentation results produced by our methodology.

4. DISCUSSION AND CONCLUSION

In this paper, we have demonstrated the suitability of TRW to mammogram mass segmentation both in terms of accuracy and computational efficiency. In particular, we have demonstrated the benefit of combining multiple potential functions based on deep learning, gaussian mixture model and shape prior. Fig. 2 demonstrates that

Table 1. Comparison between the proposed and several state-of-the-art methods.

Method	Rep.	#Images	Dataset	Dice	Run.Time
Proposed	yes	158	DDSM-BCRP	0.89	0.1s
Dhungel et al. [18]	yes	158	DDSM-BCRP	0.87	0.8s
Beller et al. [19]	yes	158	DDSM-BCRP	0.70	?
Ball et al. [4]	no	60	DDSM	0.85	?
Rahmati et al. [2]	no	100	DDSM	0.93	?
Yuan et al. [20]	no	483	DDSM	0.78	4.7s
Proposed	yes	116	INbreast	0.89	0.1s
Cardoso et al. [17]	yes	116	INbreast	0.88	?
Dhungel et al. [18]	yes	116	INbreast	0.88	0.8s

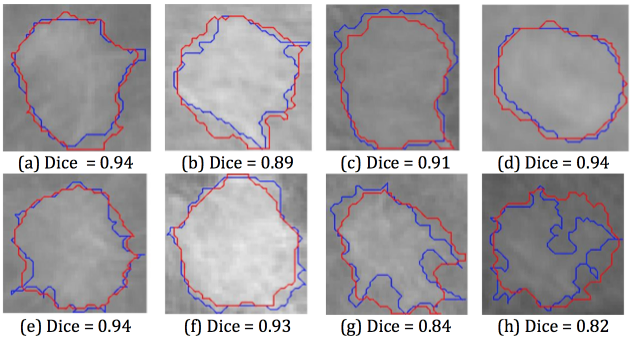

Fig. 3. Mass segmentation produced by our approach, where the red contour denotes the result of our methodology and blue represents the ground truth.

segmentation accuracy improves with the use of all potential functions together with the TRW which provides an increase in accuracy. Our method also shows good generalization capability given the small difference between the training and testing results. Furthermore, note that the pre-processing stage provides increase in accuracy. The comparison against the state-of-the-art in Tab. 1 shows that our approach is the most accurate and most efficient in the field on Inbreast and DDSM-BCRP. If one considers other subsets and annotations of DDSM, our method is still competitive, presenting the second best overall result, with [2] apparently being the most accurate. However, this is not a fair comparison because we do not have access to the annotations used in their experiment. As depicted in Fig. 3 (particularly in images (g) and (h), showing poor dice results), the main problem affecting the results of our methodology is that the produced mass segmentation tend to be a smoothed version of the ground truth annotation. This happens because of the high weight that the learning mechanism places in the mean shape term (14), which is a consequence of the small training set (available from the DDSM-BCRP and INbreast databases) that do not fully represent all possible appearance and shape variations of breast masses. We believe that by training our methodology with larger and more representative training set, we can improve segmentation accuracy. Therefore we plan to acquire such training sets and make them available for the community.

5. REFERENCES

[1] A. Jemal et al., "Cancer statistics, 2008," *CA: a cancer journal for clinicians*, vol. 58, no. 2, pp. 71–96, 2008.

[2] P. Rahmati, A. Adler et al., "Mammography segmentation with maximum likelihood active contours," *MIA*, vol. 16, no. 6, pp. 1167–1186, 2012.

[3] J. Elmore et al., "Variability in interpretive performance at screening mammography and radiologists characteristics associated with accuracy1," *Radiology*, vol. 253, no. 3, pp. 641–651, 2009.

[4] J. Ball and L. Bruce, "Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation," in *EMBS 2007*, 2007, pp. 4973–4978.

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

[6] Y. Boykov et al., "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.

[7] V. Kolmogorov et al., "Comparison of energy minimization algorithms for highly connected graphs," in *Computer Vision–ECCV 2006*, pp. 1–15. Springer, 2006.

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

[9] C. Farabet et al., "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[10] I. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. Cardoso, and J. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[11] M. Heath, K. Bowyer et al., "The digital database for screening mammography," in *International workshop on digital mammography*, 2000, pp. 212–218.

[12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[13] M. Wainwright, T. Jaakkola, and A. Willsky, "Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching," in *Workshop on Artificial Intelligence and Statistics*, 2003, vol. 21, p. 97.

[14] T. Meltzer, A. Globerson, and Y. Weiss, "Convergent message passing algorithms: a unifying view," in *UAI*, 2009, pp. 393–401.

[15] A. Dempster et al., "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B*, pp. 1–38, 1977.

[16] A. Horsch, et al., "Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies," *International journal of computer assisted radiology and surgery*, vol. 6, no. 6, pp. 749–767, 2011.

[17] J. Cardoso et al., "Closed shortest path in the original coordinates with an application to breast cancer," *International Journal of Pattern Recognition and Artificial Intelligence*, 2014.

[18] N. Dhungel, G. Carneiro, and A. Bradley, "Deep strucutred learning for mass segmentation from Mammograms," *arXiv:1410.7454 [cs.CV]*, 2014.

[19] M. Beller et al., "An example-based system to support the segmentation of stellate lesions," in *Bildverarbeitung für die Medizin 2005*, pp. 475–479. Springer, 2005.

[20] Y. Yuan et al., "A dual-stage method for lesion segmentation on digital mammograms," *Medical physics*, vol. 34, pp. 4180, 2007.