Deep Learning and Structured Prediction for the Segmentation of Mass in 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 explore the use of deep convolution and deep belief networks as potential functions in structured prediction models for the segmentation of breast masses from mammograms. In particular, the structured prediction models are estimated with loss minimization parameter learning algorithms, representing: a) conditional random field (CRF), and b) structured support vector machine (SSVM). For the CRF model, we use the inference algorithm based on tree re-weighted belief propagation with truncated fitting training, and for the SSVM model the inference is based on graph cuts with maximum margin training. We show empirically the importance of deep learning methods in producing state-of-the-art results for both structured prediction models. In addition, we show that our methods produce results that can be considered the best results to date on DDSM-BCRP and INbreast databases. Finally, we show that the CRF model is significantly faster than SSVM, both in terms of inference and training time, which suggests an advantage of CRF models when combined with deep learning potential functions.

Keywords: Deep learning, Structured output learning, Mammogram segmentation

1 Introduction

Screening mammogram is one of the most effective imaging modalitites to detect breast cancer, and it is used for the segmentation of breast masses (among other tasks), which is a challenging task due to the variable shape/size of masses [1] and their low signal-to-noise ratio (see Fig. 1). In clinical practice, lesion segmentation is usually a manual process, and so its efficacy is associated with the radiologist's expertise and workload [2], where a clear trade-off can be noted between sensitivity (Se) and specificity (Sp) in manual interpretation, with a median Se of 83.8% and Sp of 91.1% [2].

The main goal of this paper is to introduce and evaluate a new methodology for segmenting masses from mammograms based on structured prediction models

[⋆] 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).

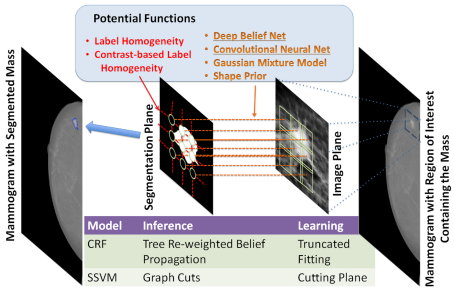


Fig. 1: Structured prediction model with a list of potential functions that include two deep learning methods and two structured prediction models

that use deep learning as their potential functions (Fig. 1). Our main contribution is the introduction of powerful deep learning appearance models, based on CNN [3, 4] and DBN [5], into the following recently proposed structured output models: a) a conditional random field (CRF), and b) structured support vector machines (SSVM). The CRF model performs inference with tree re-weighted belief propagation [6] and learning with truncated fitting [7], while the SSVM model uses graph cuts [8] for inference and cutting plane [9, 10] for training. We show that both structured output models produce comparable segmentation results, which are marginally superior to other recently proposed methods in the field in public datasets, and we also show that the use of both deep learning models is essential to reach such accurate results. Finally, we also demonstrate that the CRF model is significantly faster in terms of training and inference time, which suggests its use as the most efficient method in the field.

2 Methodology

Let us assume that the model parameter is denoted by $\mathbf{w}$, the image of the region of interest (ROI) containing the mass is denoted by $\mathbf{x} : \Omega \rightarrow \mathbb{R}$ ($\Omega \in \mathbb{R}^2$ denotes the image lattice of size $M \times M$) , the labeling is represented by $\mathbf{y} : \Omega \rightarrow \{-1, +1\}$, the training set is referred to as $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$, and the graph that links the image and labels is defined with $\mathcal{V}$ nodes and $\mathcal{E}$ edges between nodes. The learning process is based on the minimization of the empirical loss [11]:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \ell(\mathbf{x}_n, \mathbf{y}_n, \mathbf{w}), \tag{1}$$

where $\ell(\mathbf{x}, \mathbf{y}, \mathbf{w})$ is a continuous and convex loss function that defines the structured model. We explore CRF and SSVM formulations for solving Eq.(1), described in Sections 2.1 and 2.2. In particular, CRF uses the loss

$$\ell(\mathbf{x}_n, \mathbf{y}_n, \mathbf{w}) = A(\mathbf{x}_n, \mathbf{w}) - E(\mathbf{y}_n, \mathbf{x}_n; \mathbf{w}), \tag{2}$$

where $A(\mathbf{x};\mathbf{w}) = \log \sum_{\mathbf{y}\in\{-1,+1\}^{M\times M}} \exp\{E(\mathbf{y},\mathbf{x};\mathbf{w})\}$ is the log-partition function that ensures normalization, and

$$E(\mathbf{y},\mathbf{x};\mathbf{w}) = \sum_{k=1}^{K}\sum_{i\in\mathcal{V}} w_{1,k}\phi^{(1,k)}(\mathbf{y}(i),\mathbf{x}) + \sum_{l=1}^{L}\sum_{i,j\in\mathcal{E}} w_{2,l}\phi^{(2,l)}(\mathbf{y}(i),\mathbf{y}(j),\mathbf{x}), \quad (3)$$

with $\phi^{(1,k)}(.,.)$ representing one of the K potential functions between label (segmentation plane in Fig. 1) and pixel (image plane in Fig. 1) nodes, $\phi^{(2,l)}(.,.,.)$ denoting one of the L potential functions on the edges between label nodes, and $\mathbf{w} = [w_{1,1},...,w_{1,K},w_{2,1},...,w_{2,L}]^\top \in \mathbb{R}^{K+L}$ with $\mathbf{y}(i)$ being the i^{th} component of vector $\mathbf{y}$. Alternatively, the SSVM minimizes the loss function

$$\ell(\mathbf{x}_n,\mathbf{y}_n,\mathbf{w}) = \max_{\mathbf{y}\in\mathcal{Y}} \left(\Delta(\mathbf{y}_n,\mathbf{y}) + E(\mathbf{y},\mathbf{x}_n;\mathbf{w}) - E(\mathbf{y}_n,\mathbf{x}_n;\mathbf{w})\right), \quad (4)$$

where $\Delta(\mathbf{y}_n,\mathbf{y})$ returns the dissimilarity between $\mathbf{y}_n$ and $\mathbf{y}$, satistfying the conditions $\Delta(\mathbf{y}_n,\mathbf{y}) \geq 0$ and $\Delta(\mathbf{y}_n,\mathbf{y}_n) = 0$.

2.1 Conditional Random Field (CRF)

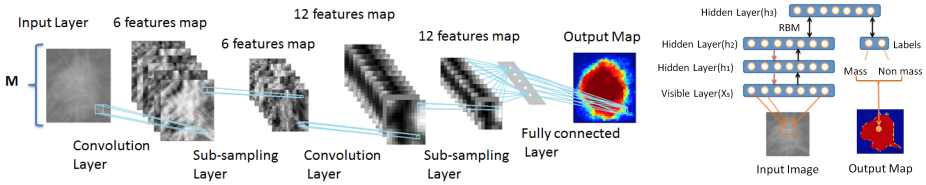
The solution of Eq.(1) using the CRF loss function in Eq.(2) involves the computation of the log-partition function $A(\mathbf{x};\mathbf{w})$. Tree re-weighted (TRW) belief propagation provides an upper bound to this log-partition function [6]:

$$A(\mathbf{x};\mathbf{w}) = \max_{\mu\in\mathcal{M}} \mathbf{w}^T\mu + H(\mu), \quad (5)$$

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 [12] (as denoted in Eq.(3), $P(\mathbf{y}|\mathbf{x},\mathbf{w}) = \exp\{E(\mathbf{y},\mathbf{x};\mathbf{w}) - A(\mathbf{x};\mathbf{w})\}$ indicating the conditional probability of the annotation $\mathbf{y}$ given the image $\mathbf{x}$ and parameters $\mathbf{w}$ (where we assume that this conditional probability function belongs to the exponential family), and $H(\mu) = -\sum_{\mathbf{y}\in\{-1,+1\}^{M\times M}} P(\mathbf{y}|\mathbf{x};\mathbf{w}) \log P(\mathbf{y}|\mathbf{x},\mathbf{w})$ is the entropy. Note that for general graphs with cycles, the marginal polytope $\mathcal{M}$ is difficult to characterize and the entropy $\mathbf{H}(\mu)$ is not tractable [7]. TRW solves these issues by first replacing the marginal polytope with a superset $\mathcal{L} \supset \mathcal{M}$ that only accounts for the local constraints of the marginals, and then approximating the entropy calculation with an upper bound [7]. The learning of $\mathbf{w}$ in (2) is achieved via gradient descent in a process called truncated fitting [7], and the inference to find the label $\mathbf{y}^*$ for an image $\mathbf{x}^*$ is based on TRW.

2.2 Structured Support Vector Machine (SSVM)

The SSVM optimization to estimate $\mathbf{w}$ consists of a regularized loss minimization problem formulated as $\mathbf{w}^* = \min_{\mathbf{w}} \|\mathbf{w}\|^2 + \lambda \sum_n \ell(\mathbf{x}_n,\mathbf{y}_n,\mathbf{w})$, with $\ell(.)$ defined



(a) CNN model (b) DBN model

Fig. 2: (a) CNN and (b) DBN models with the given mass patch as input.

in Eq.(4), where the introduction of slack variable leads to [9, 10]:

$$\begin{aligned} &\text{minimize}_{\mathbf{w}} \ \tfrac{1}{2}\|\mathbf{w}\|^2 + \tfrac{C}{N}\textstyle\sum_n \xi_n \\ &\text{subject to} \ \ E(\mathbf{y}_n, \mathbf{x}_n; \mathbf{w}) - E(\hat{\mathbf{y}}_n, \mathbf{x}_n; \mathbf{w}) \geq \Delta(\mathbf{y}_n, \hat{\mathbf{y}}_n) - \xi_n, \forall \hat{\mathbf{y}}_n \neq \mathbf{y}_n \\ &\qquad\qquad \xi_n \geq 0. \end{aligned} \tag{6}$$

In order to keep the number of constraints manageable in Eq.(6), we use the cutting plane for solving the maximization problem:

$$\hat{\mathbf{y}}_n = \arg\max_{\mathbf{y}} \Delta(\mathbf{y}_n, \mathbf{y}) + E(\mathbf{y}, \mathbf{x}_n; \mathbf{w}) - E(\mathbf{y}_n, \mathbf{x}_n; \mathbf{w}) - \xi_n, \tag{7}$$

which finds the most violated constraint for the n^{th} training sample given the parameter $\mathbf{w}$,

where $\Delta(.)$ denotes the Hamming distance [13]. The label inference for a test mammogram $\mathbf{x}$, given the learned parameters $\mathbf{w}$ from Eq.(6), is based on $\mathbf{y}^* = \arg\max_{\mathbf{y}} E(\mathbf{y}, \mathbf{x}; \mathbf{w})$, which can be optimally solved with graph cuts [8].

2.3 Potential Functions

The model in Eq.(3) can incorporate a large number of unary and binary potential functions. We propose the use of CNN and DBN in addition to the more common Gaussian mixture model (GMM) and shape prior between the nodes of image and segmentation planes (Fig. 1).

The **CNN potential function** is defined by [4] (Fig. 2-(a)):

$$\phi^{(1,1)}(\mathbf{y}(i), \mathbf{x}) = -\log P_c(\mathbf{y}(i)|\mathbf{x}, \theta_c), \tag{8}$$

where $P_c(\mathbf{y}(i)|\mathbf{x}, \theta)$ denotes the probability of labeling the pixel $i \in M \times M$ with mass or background (given the input image $\mathbf{x}$ for the ROI of the mass) and θ_c denotes the CNN parameters. A CNN model consists of multiple processing stages, each containing a convolutional layer (where the learned filters are applied to the image) and a non-linear subsampling layer that reduces the input image size for the next stage (Fig. 2), and a final stage consisting of few fully connected layers. The convolution stages compute the output at location j from input at i using the learned filter (at q^{th} stage) $\mathbf{k}^q$ and bias b^q with

$\mathbf{x}(j)^q = \sigma(\sum_{i \in M_j} \mathbf{x}(i)^{q-1} * \mathbf{k}_{ij}^q + b_j^q)$, where $\sigma(.)$ is the logistic function and M_j is the input region location; while the non-linear subsampling layers calculate subsampled data with $\mathbf{x}(j)^q = \downarrow (\mathbf{x}_j^{q-1})$, where $\downarrow (.)$ denotes a subsampling function that pools (using either the mean or max functions) the values from a region from the input data. The final stage consists of the convolution equation above using a separate filter for each output location, using the whole input from the previous layer. Inference is simply the application of this process in a feed-forward manner, and training is carried out with backpropagation to minimize the segmentation error over the training set [3, 4].

The **DBN potential function** is defined as [5] (Fig. 2-(b)):

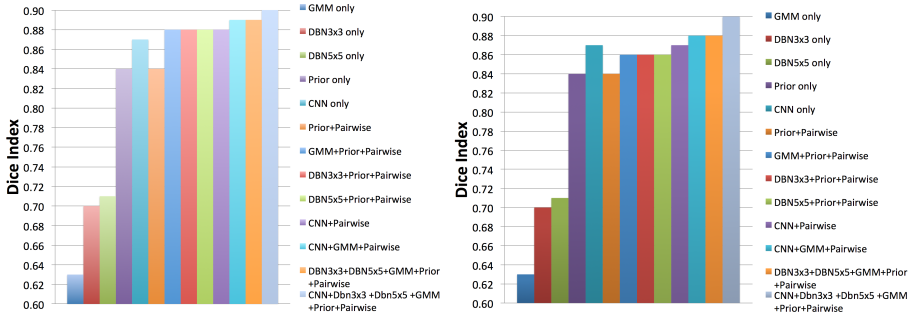
$$\phi^{(1,2)}(\mathbf{y}(i), \mathbf{x}) = -\log P_d(\mathbf{y}(i)|\mathbf{x}_S(i), \theta_{d,S}), \tag{9}$$

where $\mathbf{x}_S(i)$ is a patch extracted around image lattice position i of size $S \times S$ pixels ($S < M$, with M being the original patch size), $\theta_{d,S}$ represents the DBN parameters. The inference is based on the mean field approximation of the values in all DBN layers, followed by the computation of free energy on the top layer [5]. The learning of the DBN parameters $\theta_{d,S}$ in Eq.(9) is achieved with an iterative layer by layer training of auto-encoders using contrastive divergence, followed by a discriminative learning using backpropagation [5]. In addition to the CNN and DBN patch-based potential functions, we also use a pixel-wise **GMM model** [13] defined by $\phi^{(1,3)}(\mathbf{y}(i), \mathbf{x}) = -\log P_g(\mathbf{y}(i)|\mathbf{x}(i), \theta_g)$, where $P(.)$ is computed from the GMM class dependent probability model, learned from the training set; and the shape prior model [13] represented by $\phi^{(1,4)}(\mathbf{y}(i), \mathbf{x}) = -\log P(\mathbf{y}(i)|\theta_p)$, which computes the probability of belonging to the mass based only on the patch position (this prior is taken from the training annotations). Finally, the **pairwise potential functions** between label nodes in Eq.(3) encode label and contrast dependent labelling homogeneity as $\phi^{(2,1)}(\mathbf{y}(i), \mathbf{y}(j), \mathbf{x})$ and $\phi^{(2,2)}(\mathbf{y}(i), \mathbf{y}(j), \mathbf{x})$ [11].

3 Experiments

3.1 Materials and Methods

The evaluation of our methodology is performed on two publicly available datasets: INbreast [14] and DDSM-BCRP [15].The INbreast dataset comprises a set of 56 cases containing 116 accurately annotated masses. We divide this dataset into mutually exclusive train and test sets, each containing 58 images. The DDSM-BCRP [15] dataset consists of 39 cases (77 annotated images) for training and 40 cases (81 annotated images) for testing. We used Dice index to assess the segmentation accuracy. Efficiency is estimated with the training and testing time of the segmentation algorithm, obtained on a standard computer (Intel(R) Core(TM) i5-2500k 3.30GHz CPU with 8GB RAM). The ROI to be segmented is obtained by a manual input of location and scale, similarly to other works in the field [13, 16, 17]. It is important to realize that the segmentation of masses from these manually labeled regions is an important step in mass classification, so it is still



(a) CRF model in Eq.(2) (b) SSVM model in Eq.(4)

Fig. 3: Dice index over the test set of INbreast of the CRF (a) and SSVM (b) models, using various subsets of the potential functions.

an open problem [18] because of the challenges involved, such as spicules segmentation, low signal-to-noise ratio, and lack of robust shape and appearance models. This ROI forms a rectangular bounding box that is resized to 40 x 40 pixels using bicubic interpolation and pre-processed with Ball and Bruce technique [1]. The model selection for the CNN/DBN structures is performed via cross validation on the training set, and for the CNN, the net structure of the first and second stages have filters of size 5×5 and a subsampling based on max pooling. The final stage of the CNN has a fully connected layer with 588 nodes and an output layer with 40×40 nodes (i.e., same size of the input layer). We use two DBN models, like the one shown in Fig. 2(b), where each of the three layers contains 50 nodes and input patches of sizes 3×3 and 5×5 (i.e., $S = 3, 5$, respectively).

3.2 Results

Fig. 3 shows the importance of adding each potential function in the model Eq.(3) to improve the Dice index using both CRF and SSVM. We show these results using several subsets of the potential functions introduced in Sec. 2.3 (i.e., the potentials $\phi^{(1,k)}$ for $k = \{1, 2, 3, 4\}$ with 3×3 and 5×5 denoting the image patch size used by the DBN). The Dice index of our methodology using all potential functions on the train set of INbreast is similar to test set at 0.93 using CRF and 0.95 using SSVM. It is also worth mentioning that the results on the INbreast test set when we do not use preprocessing [1] falls to 0.85 using all potential functions for both models (similar results are obtained on DDSM).

Tab. 1 shows the Dice index and average training (for the whole training set) and testing times (per image) of our approach with all potential functions (CNN+DBN3x3 + DBN5x5 + GMM + Prior + Pairwise) using the CRF and SSVM models on DDSM-BCRP and INbreast test sets.

Table 1: Comparison of the proposed and state-of-the-art methods on test sets.

Method	#Images	Dataset	Dice Index	Test Run.Time	Train Run. time
Proposed CRF model	116	INbreast	0.90(0.06)	0.1s	360s
Proposed SSVM model	116	INbreast	0.90(0.06)	0.8s	1800s
Cardoso et al. [16]	116	INbreast	0.88	?	?
Dhungel et al. [13]	116	INbreast	0.88	0.8s	?
Proposed CRF model	158	DDSM-BCRP	0.90(0.06)	0.1s	383s
Proposed SSVM model	158	DDSM-BCRP	0.90(0.06)	0.8s	2140s
Dhungel et al. [13]	158	DDSM-BCRP	0.87	0.8s	?
Beller et al. [17]	158	DDSM-BCRP	0.70	?	?

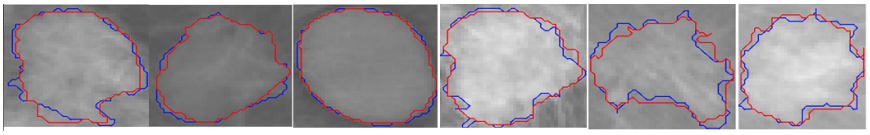


Fig. 4: Segmentation results by the CRF model on INbreast where the blue denotes the manual annotation and red denotes automatic segmentation.

4 Discussion and Conclusions

From the results in Fig. 3, we notice that the use of deep learning based potential functions provide a significant improvement when compared with the shape prior alone. Also, the combination of GMM and deep learning models improve both the CRF and SSVM models. Another important observation is the fact that the image preprocessing [1] is important empirically. The comparison with other methods in Table 1 shows that our methodology produces the best results(0.90 v 0.88 and 0.90 v 0.87) for both databases. Our CRF and SSVM models demonstrate equivalent results (0.90) on both data sets. However, assuming a standard deviation of 0.06, a t-test indicates that our methods perform significantly ($p<0.01$) better than the previous methods [13, 16, 17]. The comparison in terms of train and test times shows a significant advantage to the CRF model over SSVM model. There are some important notes to make about the training and testing processes in these results: 1) we tried different CNN structures and the combination of more than one CNN model as additional potential functions, but the single CNN model detailed in Sec. 3.1 produced the best cross validation results; 2) for the DBN models, we tried different input sizes (3×3 and 7×7 patches), but the combinations of the ones detailed in Sec. 3.1 provided the best cross-validation results; and 3) both CRF and SSVM models estimate a much larger weight to the CNN potential function compared to others in Sec. 2.3, indicating that this is the most important potential function, but the CNN model alone overfits the training data (with a Dice of 0.87 on test and 0.95 on training), so the structural prediction models (CRF and SSVM) serve as a regularizer to the CNN model. Finally, from the visual results in Fig. 4, our CRF model produces quite accurate segmentation results even in the presence of moderately sharp corners and cusps.

References

1. Ball, J., Bruce, L.: Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation. In: EMBS 2007. 29th Annual International Conference of the IEEE, IEEE (2007) 4973–4978
2. Elmore, J.G., Jackson, S.L., Abraham, L., et al.: Variability in interpretive performance at screening mammography and radiologists characteristics associated with accuracy1. Radiology **253**(3) (2009) 641–651
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. Volume 1. (2012) 4
4. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361** (1995)
5. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786) (2006) 504–507
6. Wainwright, M.J., Jaakkola, T.S., Willsky, A.S.: Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching. In: Workshop on Artificial Intelligence and Statistics. Volume 21., Society for Artificial Intelligence and Statistics Np (2003) 97
7. Domke, J.: Learning graphical model parameters with approximate marginal inference. arXiv preprint arXiv:1301.3193 (2013)
8. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI **23**(11) (2001) 1222–1239
9. Szummer, M., Kohli, P., Hoiem, D.: Learning crfs using graph cuts. In: Computer Vision–ECCV 2008. Springer (2008) 582–595
10. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. In: JMLR. (2005) 1453–1484
11. Nowozin, S., Lampert, C.: Structured learning and prediction in computer vision. Foundations and Trends in Computer Graphics and Vision **6**(3–4) (2011) 185–365
12. Meltzer, T., Globerson, A., Weiss, Y.: Convergent message passing algorithms: a unifying view. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press (2009) 393–401
13. Dhungel, N., Carneiro, G., Bradley, A.P.: Deep structured learning for mass segmentation from mammograms. arXiv preprint arXiv:1410.7454 (2014)
14. Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Academic Radiology **19**(2) (2012) 236–248
15. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: The digital database for screening mammography. In: Proceedings of the 5th international workshop on digital mammography. (2000) 212–218
16. Cardoso, J.S., Domingues, I., Oliveira, H.P.: Closed shortest path in the original coordinates with an application to breast cancer. International Journal of Pattern Recognition and Artificial Intelligence (2014)
17. Beller, M., Stotzka, R., Müller, T.O., Gemmeke, H.: An example-based system to support the segmentation of stellate lesions. In: Bildverarbeitung für die Medizin 2005. Springer (2005) 475–479
18. Rahmati, P., Adler, A., Hamarneh, G.: Mammography segmentation with maximum likelihood active contours. Medical image analysis **16**(6) (2012) 1167–1186