Combining Deep Learning and Structured Prediction for Segmenting Masses in Mammograms

Neeraj Dhungel, Gustavo Carneiro and Andrew P. Bradley

Abstract The segmentation of masses from mammogram is a challenging problem because of their variability in terms of shape, appearance and size, and the low signal to noise ratio of their appearance. We address this problem with structured output prediction models that use potential functions based on deep convolution neural network (CNN) and deep belief network (DBN). The two types of structured output prediction models that we study in this work are the conditional random field (CRF) and structured support vector machines (SSVM). The label inference for CRF is based on tree re-weighted belief propagation (TRW) and training is achieved with the truncated fitting algorithm; whilst for the SSVM model, inference is based upon graph cuts and training depends on a max-margin optimisation. We compare the results produced by our proposed models using the publicly available mammogram datasets DDSM-BCRP and INbreast, where the main conclusion is that both models produce results of similar accuracy, but the CRF model shows faster training and inference. Finally, when compared to the current state of the art in both datasets, the proposed CRF and SSVM models show superior segmentation accuracy.

Neeraj Dhungel
Australian Centre for Visual Technologies, The University of Adelaide, Adelaide,Australia e-mail: neeraj.dhungel@adelaide.edu.au

Gustavo Carneiro
Australian Centre for Visual Technologies, The University of Adelaide, Adelaide, Australia e-mail: gustavo.carneiro@adelaide.edu.au

Andrew P. Bradley
School of Information Technology and Electrical Engineering, The University of Queensland, Queensland, Australia e-mail: a.bradley@itee.uq.edu.au

This work is an extension of the paper published by the same authors at the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015) [1].

1 Introduction

Statistical findings published by World Health Organisation (WHO) [2] reveal that 23% of all diagnosed cancers and 14% of all cancer related deaths among women are due to breast cancer. These numbers show that breast cancer is one of the major diseases affecting the lives of many women across the globe. One of the keys to reduce these number is the early detection of this disease, which is task that is mostly based on mammography screening. An important activity involved in this screening process is the detection and classification of breast masses, which is difficult because of the variable size, shape and appearance of masses [3] and their low signal-to-noise ratio (see Fig. 1). In this work we focus on the problem of accurate mass segmentation because we assume that such precise segmentation is important for the sub-sequent mass classification task [4, 5]. In clinical practice, the task of detecting and segmenting masses from mammograms typically consists of a manual process performed by radiologists. This process can introduce variability depending on the radiologist's expertise and the number of mammograms to be analysed at one sitting, which can reduce the efficacy of the screening process. In a recent study [6], it has been shown that there is a clear trade-off between sensitivity (Se) and specificity (Sp) in manual interpretation, with a median Se of 84% and Sp of 91%.

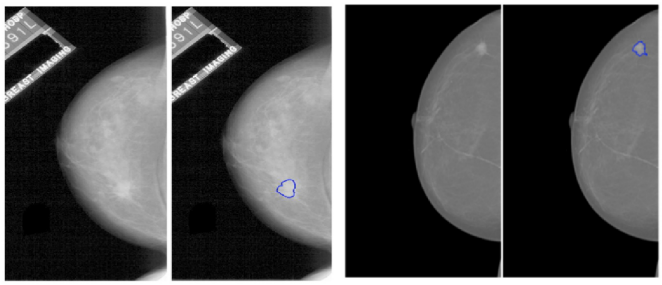


Fig. 1 Examples from INbreast [7] and DDSM-BRCP [8] databases with blue contour denoting the mass lesion with the blue contour.

Regardless of the development of numerous breast mass segmentation techniques, computer-aided diagnosis (CAD) systems, which depend on accurate breast mass segmentation methods, are not widely used in clinical practice. In fact, it has been observed that the use of CAD systems can reduce screening accuracy by increasing the rate of biopsies without improving the detection of invasive breast cancer [9]. We believe that one of the reasons is the lack of an easily reproducible and reliable assessment mechanism that provides a clear comparison between competing methodologies, which can lead to a better informed decision process related to the selection of appropriate algorithms for CAD systems. We have addressed this issue

in previous versions of this work [10, 11], where we propose quantitatively comparison mechanisms on the publicly available databases DDSM-BCRP [8] and INbreast dataset [7]. Another reason for the relatively poor performance of most of the currently available breast mass segmentation methods lies in their reliance on more traditional image processing and segmentation techniques, such as active contours, which typically produce sub-optimal results due to their non-convex cost functions. Differently from these methods, our approach is based on a machine learning technique that estimates optimal models directly from annotated data, and for this reason our approach has the potential to deliver improved segmentation accuracy, a result previously demonstrated in other medical image analysis problems [12].

In this work, we propose a new approach for segmenting breast masses from mammograms using two types of structured output prediction models: 1) conditional random field (CRF) [11, 13] and 2) structural support vector machine (SSVM) [10, 14]. Our main contribution is related to the introduction of powerful deep learning networks into the CRF and SSVM models above, based on the deep convolutional neural network (CNN) [15, 16] and the deep neural network (DBN) [17]. These deep learning architectures are able to extract image features in a fully automated manner, instead of being hand-crafted. In addition, these CNNs and DBNs have produced state-of-the-art results in several computer vision problems [15, 18], and we believe that these methodologies have the potential to produce competitive results in mass segmentation from mammography. The CRF model uses tree re-weighted belief propagation [19] for inference and truncated fitting for training [13], whilst SSVM performs label inference with graph cuts [20] and the parameters learning with the cutting plane algorithm [21, 14]. Given that these training algorithms learn all parameters for the structured output prediction models using the manually annotated training 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 modelling in a robust manner the shape and appearance variations of masses encountered in the training data if enough annotated training data is available. We test our proposed methodologies on the publicly available datasets INbreast [7] and DDSM-BCRP [8], and our methodologies produce state-of-the-art results in terms of accuracy and running time. Moreover, comparing the CRF and SSVM models, we note that they produce comparable results in terms of segmentation accuracy, but the CRF model is more efficient in terms of training and testing.

2 Literature Review

Currently, the majority of the methodologies developed for the problem of segmenting masses from mammograms is based on statistical thresholding, dynamic programming models, morphological operators and active contour models. A statistical thresholding method that distinguishes pixels inside the mass area from those outside has been developed by Catarious et al. [22]. Although relatively successful, the main drawback of this type of approach is that it is not robust to low contrast im-

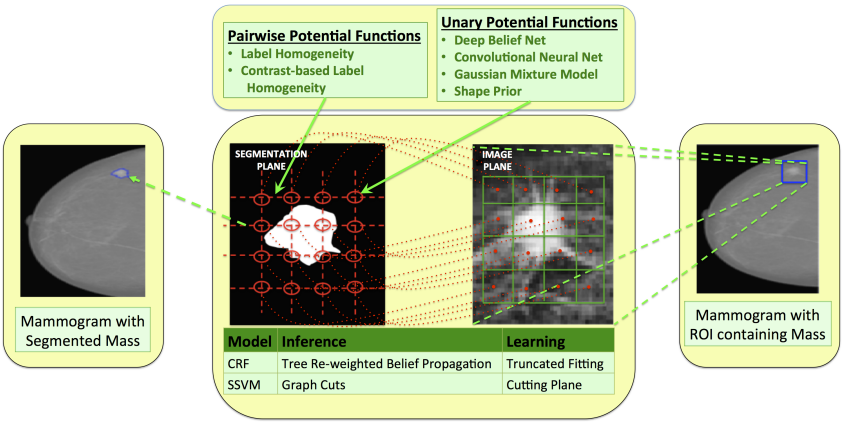


Fig. 2 The proposed structured output prediction models with a list of unary and pairwise potential functions for mass segmentation in mammograms, including the deep learning networks.

ages [4]. Song et al. [23] have extended this model with a statistical classifier based on edge gradient, pixel intensity and shape characteristics, where the segmentation is found by estimating the minimum cut of a graph representation of the image using dynamic programming. Similar dynamic programming models have also been applied by Timp et al. [24], Dominguez et al. [25] and Yu et al. [26]. These approaches are similar to our proposed structured output prediction models, with the exception that they do not use structured learning to estimate the weights of the potential functions, which generally leads to sub-optimal performance. Morphological operators, such as the watershed method [27] or region growing [5], have also been explored for the mass segmentation problem, but these operators have been shown to be rather limited in providing sufficiently accurate results mainly because they only explore semi-local grey level distributions without considering higher level information (e.g., shape model).

Active contour models are probably the most explored methodology for breast mass segmentation. The most accurate model reported in the field is the one proposed by Rahmati et al. [4], which is a level set method based on the maximum likelihood segmentation without edges that is particularly well adapted to noisy images with weak boundaries. Several other papers also propose mass segmentation methods based on standard active contour models [28, 29, 30, 31, 32]. The major drawback of active contour models lies in their need of a good initialisation for the inference process due to the usual non-convexity of the energy function. Moreover, the weights of the terms forming the energy function of the active contour models models are usually arbitrarily defined, or estimated via a cross-validation process that generally do not produce an optimal estimation of these weights.

3 Methodology

We start this section with an explanation of the learning process of our structured output prediction model [33]. Assume that we have an annotated dataset $\mathscr{D}$ containing images of the region of interest (ROI) of the mass, represented by $\mathbf{x} : \Omega \rightarrow \mathbf{R}$ ($\Omega \in \mathbf{R}^2$), and the respective manually provided segmentation mask $\mathbf{y} : \Omega \rightarrow \{-1,+1\}$, where $\mathscr{D} = (\mathbf{x},\mathbf{y})_{i=1}^{|\mathscr{D}|}$. Also assume that the parameter of our structured output prediction model is denoted by θ and the graph $\mathscr{G} = (\mathscr{V},\mathscr{E})$ links the image $\mathbf{x}$ and labels $\mathbf{y}$, where $\mathscr{V}$ represents the set of graph nodes and $\mathscr{E}$, the set of edges. The process of learning the parameter of our structured prediction model is done via the minimisation of the following empirical loss function [33]:

$$\theta^* = \arg\min_{\theta} \frac{1}{|\mathscr{D}|} \sum_{i=1}^{|\mathscr{D}|} \ell(\mathbf{x}_i, \mathbf{y}_i, \theta), \tag{1}$$

where $\ell(\mathbf{x},\mathbf{y},\theta)$ is a continuous and convex loss function being minimized that defines the structured model. We use CRF and SSVM formulations for solving (1), which are explained in detail in Sec. 3.1 and 3.2, respectively, and we explain the potential functions used for both models in Sec. 3.3. In particular, the CRF formulation uses the loss

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

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

$$E(\mathbf{y},\mathbf{x};\theta) = \sum_{k=1}^{K} \sum_{i \in \mathscr{V}} \theta_{1,k} \psi^{(1,k)}(\mathbf{y}(i),\mathbf{x}) + \sum_{l=1}^{L} \sum_{i,j \in \mathscr{E}} \theta_{2,l} \psi^{(2,l)}(\mathbf{y}(i),\mathbf{y}(j),\mathbf{x}), \tag{3}$$

In (3), $\psi^{(1,k)}(.,.)$ denotes one of the K potential functions between label (segmentation plane in Fig. 2) and pixel (image plane in Fig. 2) nodes, $\psi^{(2,l)}(.,.,.)$ denoting one of the L potential functions on the edges between label nodes, $\theta = [\theta_{1,1},...,\theta_{1,K},\theta_{2,1},...,\theta_{2,L}]^\top \in \mathbf{R}^{K+L}$, and $\mathbf{y}(i)$ being the i^{th} component of vector $\mathbf{y}$. Similarly, the SSVM uses the following loss function

$$\ell(\mathbf{x}_i, \mathbf{y}_i, \theta) = \max_{\mathbf{y} \in \mathscr{Y}} \left(\Delta(\mathbf{y}_i, \mathbf{y}) + E(\mathbf{y}, \mathbf{x}_i; \theta) - E(\mathbf{y}_i, \mathbf{x}_i; \theta) \right), \tag{4}$$

where $\Delta(\mathbf{y}_i,\mathbf{y})$ represents the dissimilarity between $\mathbf{y}_i$ and $\mathbf{y}$, which satistfies the conditions $\Delta(\mathbf{y}_i,\mathbf{y}) \geq 0$ for $\mathbf{y}_i \neq \mathbf{y}$ and $\Delta(\mathbf{y}_i,\mathbf{y}_i) = 0$.

3.1 Conditional Random Field (CRF)

The solution of (1) using the CRF loss function in (2) involves the computation of the log-partition function $A(\mathbf{x};\theta)$. The tree re-weighted belief propagation algorithm provides the following upper bound to this log-partition function [19]:

$$A(\mathbf{x};\theta) = \max_{\mu \in \mathscr{M}} \theta^T \mu + H(\mu), \tag{5}$$

where $\mathscr{M} = \{\mu' : \exists \theta, \mu' = \mu\}$ denotes the marginal polytope, $\mu = \sum_{\mathbf{y} \in \{-1,+1\}^{|\Omega| \times |\Omega|}} P(\mathbf{y}|\mathbf{x},\theta) f(\mathbf{y})$, with $f(\mathbf{y})$ denoting the set of indicator functions of possible configurations of each clique and variable in the graph [34] (as denoted in (3)), $P(\mathbf{y}|\mathbf{x},\theta) = \exp\{E(\mathbf{y},\mathbf{x};\theta) - A(\mathbf{x};\theta)\}$ indicating the conditional probability of the annotation $\mathbf{y}$ given the image $\mathbf{x}$ and parameters θ (where we assume that this conditional probability function belongs to the exponential family), and $H(\mu) = -\sum_{\mathbf{y} \in \{-1,+1\}^{|\Omega| \times |\Omega|}} P(\mathbf{y}|\mathbf{x};\theta) \log P(\mathbf{y}|\mathbf{x},\theta)$ is the entropy. Note that for general graphs with cycles (such as the case in this paper), the marginal polytope $\mathscr{M}$ is difficult to characterise and the entropy $\mathbf{H}(\mu)$ is not tractable [13]. Tree re-weighted belief propagation (TRW) solves these issues by first replacing the marginal polytope with a superset $\mathscr{L} \supset \mathscr{M}$ that only accounts for the local constraints of the marginals, and then approximating the entropy calculation with an upper bound. Specifically,

$$\mathscr{L} = \{\mu : \sum_{\mathbf{y}(c) \setminus \mathbf{y}(i)} \mu(\mathbf{y}(c)) = \mu(\mathbf{y}(i)), \sum_{\mathbf{y}(i)} \mu(\mathbf{y}(i)) = 1\} \tag{6}$$

replaces $\mathscr{M}$ in (5) and represents the local polytope (with $\mu(\mathbf{y}(i)) = \sum_{\mathbf{y}'} P(\mathbf{y}'|\mathbf{x},\theta)$ $\delta(\mathbf{y}'(i) - \mathbf{y}(i))$ and $\delta(.)$ denoting the Dirac delta function), c indexes a graph clique, and the entropy approximation (that replaces $H(\mu)$ in (5)) is defined by

$$\widetilde{H}(\mu) = \sum_{\mathbf{y}(i)} H(\mu(\mathbf{y}(i))) - \sum_{\mathbf{y}(c)} \rho_c I(\mu(\mathbf{y}(c)), \tag{7}$$

where $H(\mu(\mathbf{y}(i))) = -\sum_{s(i)} \mu(\mathbf{y}(i)) \log \mu(\mathbf{y}(i))$ 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 providing the upper bound on the entropy. Therefore, the estimation of $A(\mathbf{x};\theta)$ and associated marginals in (5) is based on the following message-passing updates [13]:

$$\begin{aligned} m_c(\mathbf{y}(i)) \propto \sum_{\mathbf{y}(c) \setminus \mathbf{y}(i)} & \exp\left\{ \frac{1}{\rho_c} \psi_c(\mathbf{y}(i), \mathbf{y}(j); \theta \right\} \\ & \prod_{j \in c \setminus i} \exp\left\{ \frac{1}{\rho_c} \psi_i(\mathbf{y}(i), \mathbf{x}; \theta) \right\} \frac{\prod_{d: j \in d} m_d(s(j))^{\rho_d}}{m_c(s(j))}, \end{aligned} \tag{8}$$

where $\phi_i(\mathbf{y}(i), \mathbf{x}; \theta) = \sum_{k=1}^{K} w_{1,k} \psi^{(1,k)}(\mathbf{y}(i), \mathbf{x})$ and $\psi_c(\mathbf{y}(i), \mathbf{y}(j); \theta) =$ $\sum_{l=1}^{L} w_{2,l} \phi^{(2,l)}(\mathbf{y}(i), \mathbf{y}(j), \mathbf{x})$ (see (3)). Once the message passing algorithm con-

verges [19], the beliefs for the associated marginals are written as:

$$\begin{aligned}\mu_c(\mathbf{y}(c)) &\propto \frac{1}{\rho_c}\psi_c(\mathbf{y}(i),\mathbf{y}(j))\prod_{i\in c}\psi_i(\mathbf{y}(i),\mathbf{x};\theta)\frac{\prod_{d:j\in d} m_d(\mathbf{y}(j))^{\rho_d}}{m_c(\mathbf{y}(i))} \\ \mu_i(\mathbf{y}_i) &\propto \exp(\psi_i(\mathbf{y}(i),\mathbf{x};\theta))\prod_{d:i\in d} m_d(\mathbf{y}(i))^{\rho_d}.\end{aligned} \tag{9}$$

The learning process involved in the estimation of θ is typically based on gradient descent that minimizes the loss in (2) and should run until convergence, which is defined by the change rate of θ between successive gradient descent iterations. However, as noted by Domke [13], there are problems with this approach, where large thresholds in this change rate can lead to bad suboptimal estimations, and tight thresholds result in slow convergence. These issues are circumvented by the truncated fitting algorithm [13], which uses a fixed number of iterations (i.e., no threshold is used in this training algorithm). We refer the reader to [13] for more details on this training algorithm.

3.2 Structured Support Vector Machine (SSVM)

The SSVM optimization to estimate θ consists of a regularized loss minimization problem formulated as $\theta^* = \min_\theta \|\theta\|^2 + \lambda \sum_i \ell(\mathbf{x}_i,\mathbf{y}_i,\theta)$, with $\ell(.)$ defined in (4). The introduction of slack variable leads to the following optimization problem [21, 14]:

$$\begin{aligned}&\text{minimize}_\theta \ \tfrac{1}{2}\|\theta\|^2 + \tfrac{C}{|\mathscr{D}|}\textstyle\sum_i \xi_i \\ &\text{subject to} \ \ E(\mathbf{y}_i,\mathbf{x}_i;\theta) - E(\hat{\mathbf{y}}_i,\mathbf{x}_i;\theta) \geq \Delta(\mathbf{y}_i,\hat{\mathbf{y}}_i) - \xi_i, \forall \hat{\mathbf{y}}_i \neq \mathbf{y}_i \\ &\qquad\quad \xi_i \geq 0.\end{aligned} \tag{10}$$

This optimization is a quadratic programming problem involving an intractably large number of constraints. In order to keep the number of constraints manageable, we use the cutting plane method that keeps a relatively small subset of the constraints by solving the maximization problem:

$$\hat{\mathbf{y}}_i = \arg\max_{\mathbf{y}} \Delta(\mathbf{y}_i,\mathbf{y}) + E(\mathbf{y},\mathbf{x}_i;\theta) - E(\mathbf{y}_i,\mathbf{x}_i;\theta) - \xi_i, \tag{11}$$

which finds the most violated constraint for the i^{th} training sample given the parameter θ. Then if the right hand side is strictly larger than zero, the most violated constraint is included in the constraint set and (10) is re-solved. This iterative process runs until no more violated inequalities are found. Note that if we remove the constants from (11), the optimization problem is simply: $\hat{\mathbf{y}}_i = \arg\max_{\mathbf{y}} \Delta(\mathbf{y}_i,\mathbf{y}) + E(\mathbf{y},\mathbf{x}_i;\theta)$, which can be efficiently solved using graph cuts [20] if the function $\Delta(.,.)$ can be properly decomposed in the label space. A simple example that works with graph cuts is $\Delta(\mathbf{y},\mathbf{y}_i) = \sum_i 1 - \delta(\mathbf{y}(i) - \mathbf{y}_i(i))$, which rep-

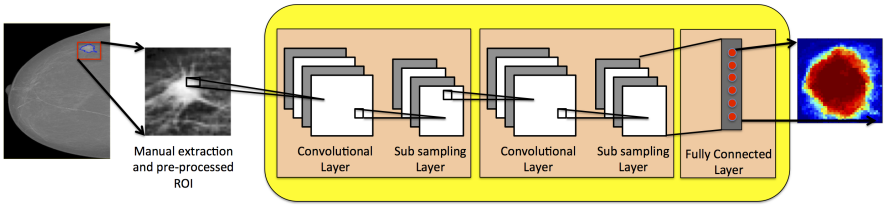


Fig. 3 CNN Model with the input $\mathbf{x}$ (mass ROI from the mammogram) and the segmentation of the whole input with $\mathbf{y}(i) \in \{-1,+1\}$, denoting the absence (blue) or presence (red) of mass, respectively, and $i \in |\Omega| \times |\Omega|$.

resents the Hamming distance and can be decomposed in the label space. Therefore, we use it in our methodology.

The label inference for a test mammogram $\mathbf{x}$, given the learned parameters θ from (10), is based on the following inference:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} E(\mathbf{y}, \mathbf{x}; \theta), \tag{12}$$

which can be efficiently and optimally solved for binary problems with graph cuts [20].

3.3 Potential Functions

It is worth noticing that the model in (3) can incorporate a large number of different types of potential functions. We propose the use of the deep convolutional neural networks (CNN) and deep belief networks (DBN), in addition to the more common Gaussian mixture model (GMM) and shape prior between the nodes of image and segmentation planes (see Fig. 2). Furthermore, we also propose the use of common pairwise potential functions.

3.3.1 CNN Potential Function

The CNN potential function is defined by [16] (Fig. 3):

$$\psi^{(1,1)}(\mathbf{y}(i), \mathbf{x}) = -\log P_{\text{CNN}}(\mathbf{y}(i)|\mathbf{x}, \theta_{\text{CNN}}), \tag{13}$$

where $P_{\text{CNN}}(\mathbf{y}(i)|\mathbf{x}, \theta_{\text{CNN}})$ denotes the probability of labeling the pixel $i \in |\Omega| \times |\Omega|$ with mass or background (given the whole input image $\mathbf{x}$ for the ROI of the mass), and θ_{CNN} denotes the CNN parameters. A CNN model consists of multiple processing stages, with each stage comprising two layers (the convolutional layer, where the learned filters are applied to the image, and the non-linear subsampling layer,

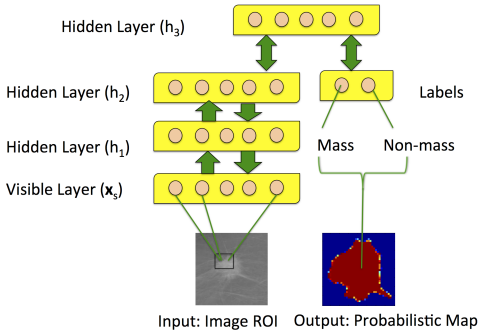


Fig. 4 DBN model with variables $\mathbf{x}$ (mass ROI from the mammogram) and classification $\mathbf{y} \in \{-1,+1\}$, denoting the absence or presence of mass, respectively.

that reduces the input image size for the next stage - see Fig. 3), and a final stage consisting of a fully connected layer. Essentially, 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 using $\mathbf{x}(j)^q = \sigma(\sum_{i \in M_j} \mathbf{x}(i)^{q-1} * \mathbf{k}_{ij}^q + b_j^q)$, where $\sigma(.)$ is the logistic function and M_j is the input region addresses; 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 stochastic gradient descent to minimize the segmentation error over the training set (via back propagation) [16].

3.3.2 DBN Potential Function

The DBN potential function is defined as [17]:

$$\psi^{(1,2)}(\mathbf{y}(i),\mathbf{x}) = -\log P_{\text{DBN}}(\mathbf{y}(i)|\mathbf{x}_S(i),\theta_{\text{DBN}}), \tag{14}$$

where $\mathbf{x}_S(i)$ is a patch extracted around image lattice position i of size $|\Omega| \times |\Omega|$ pixels, θ_{DBN} represents the DBN parameters (below, we drop the dependence on θ_{DBN} for notation simplicity), and

$$P_{\text{DBN}}(\mathbf{y}(i)|\mathbf{x}_S(i)) \propto \sum_{\mathbf{h}_1} ... \sum_{\mathbf{h}_Q} P(\mathbf{x}_S(i),\mathbf{y}(i),\mathbf{h}_1,...,\mathbf{h}_Q), \tag{15}$$

with the DBN model consisting of a network with Q layers denoted by:

$$P(\mathbf{x}_S(i),\mathbf{y}(i),\mathbf{h}_1,...,\mathbf{h}_Q) = \quad P(\mathbf{h}_Q,\mathbf{h}_{Q-1},\mathbf{y}(i))\left(\prod_{q=1}^{Q-2} P(\mathbf{h}_{q+1}|\mathbf{h}_q)\right) P(\mathbf{h}_1|\mathbf{x}_S(i)), \tag{16}$$

where $\mathbf{h}_q \in \mathbf{R}^{|q|}$ represents the hidden variables at layer q containing $|q|$ nodes. The first term in (16) is defined by:

$$-\log(P(\mathbf{h}_Q,\mathbf{h}_{Q-1},\mathbf{y}(i))) \propto -\mathbf{b}_Q^\top \mathbf{h}_Q - \mathbf{a}_{Q-1}^\top \mathbf{h}_{Q-1} - \mathbf{a}_s^\top \mathbf{y}(i) - \mathbf{h}_Q^\top \mathbf{W}\mathbf{h}_{Q-1} - \mathbf{h}_Q^\top \mathbf{W}_s \mathbf{y}(i), \tag{17}$$

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 function $\sigma(.)$, which means that $P(\mathbf{h}_{q+1}(i)|\mathbf{h}_q) = \sigma(\mathbf{b}_{q+1}(i) + \mathbf{W}_i \mathbf{h}_q)$. The inference is based on 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 [17]. The learning of the DBN parameters θ_{DBN} in (18) is achieved with an iterative layer by layer training of auto-encoders using contrastive divergence [17].

3.3.3 GMM Potential Function

The GMM potential function is defined by:

$$\psi^{(1,3)}(\mathbf{y}(i),\mathbf{x}) = -\log P_{\text{GMM}}(\mathbf{y}(i)|\mathbf{x}(i),\theta_{\text{GMM}}), \tag{18}$$

where $P_{\text{GMM}}(\mathbf{y}(i)|\mathbf{x}(i),\theta_{\text{GMM}}) = (1/Z)\sum_{m=1}^{G} \pi_m \mathcal{N}(\mathbf{x}(i);\mathbf{y}(i),\mu_m,\sigma_m) P(\mathbf{y}(i))$ with $\theta_{\text{GMM}} = [\pi_m,\mu_m,\sigma_m]_{m=1}^{G}$, $\mathcal{N}(.)$ 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_{GMM} in (14) is learned with the expectation-maximization (EM) algorithm [35] using the annotated training set.

3.3.4 Shape Prior Potential Function

The shape prior potential function is computed from the average annotation (estimated from the training set) at each image lattice position $i \in \Omega$, as follows:

$$\psi^{(1,4)}(\mathbf{y}(i),\mathbf{x}) = -\log P_{\text{prior}}(\mathbf{y}(i)|\theta_{\text{prior}}), \tag{19}$$

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

3.3.5 Pairwise Potential Functions

The pairwise potential functions between label nodes in (3) encode label and contrast dependent labelling homogeneity. In particular, the label homogeneity is defined by:

$$\psi^{(2,1)}(\mathbf{y}(i),\mathbf{y}(j),\mathbf{x}) = 1-\delta(\mathbf{y}(i)-\mathbf{y}(j)), \tag{20}$$

and the contrast dependent labelling homogeneity that we use is as follows [21]:

$$\psi^{(2,2)}(\mathbf{y}(i),\mathbf{y}(j),\mathbf{x}) = (1-\delta(\mathbf{y}(i)-\mathbf{y}(j)))C(\mathbf{x}(i)-\mathbf{x}(j)) \tag{21}$$

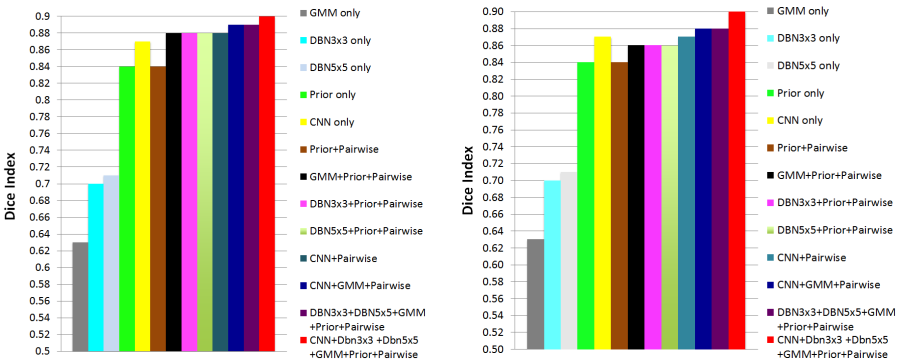
where $C(\mathbf{x}(i),\mathbf{x}(j)) = e^{-(\mathbf{x}(i)-\mathbf{x}(j))^2}$.

4 Experiments

In this section, we first introduce the datasets used, followed by an explanation of the experimental setup and the results achieved.

4.1 Materials and Methods

We assess performance of our methodology on two publicly available datasets: INbreast [7] and DDSM-BCRP [8]. The INbreast [7] dataset consist of set of 56 cases containing 116 accurately annotated masses. We divide this dataset into mutually exclusive training and testing sets, each containing 28 cases (58 annotated images each). The DDSM-BCRP [8] dataset consists of 39 cases (77 annotated images) for training and 40 cases (81 annotated images) for testing. 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. The ROI to be segmented is obtained by extracting a rectangular bounding box from around the centre of the manual annotation, where the size for each dimension of the rectangle is produced by the size of the annotation plus two pixels [36]. We use the preprocessing method by Ball and Bruce [28] in order to increase the contrast of the input image. This ROI is then resized to 40 x 40 pixels using bicubic interpolation. The model selection process for the structure of the CNN and DBN is performed via cross validation on the training set, and for the CNN, the net structure is the one in Fig. 3, where the first stage has 6 filters of size 5×5 and the second stage has 12 filters of size 5×5, and the subsampling method after each of these stages uses max pooling that reduces the input to half of its initial size in both stages. The final stage of the CNN has a fully connected layer with 588 nodes and an output layer with 1600 nodes which is reshaped to 40×40 nodes (i.e., same size of the input layer). For the DBN, the model is the one shown in Fig. 4 with $\mathbf{h}_1$, $\mathbf{h}_2$ and $\mathbf{h}_3$ each containing 50 nodes, with input patches



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

Fig. 5 Dice index on the test set of INbreast dataset for our CRF (a) and SSVM (b) models, using various subsets of the unary and pairwise potential functions.

of sizes 3×3 and 5×5. We assessed the efficiency of our segmentation methodology with the mean execution time per image on a computer with the following configuration: Intel(R) Core(TM) i5-2500k 3.30GHz CPU with 8GB RAM.

4.2 Results

The experimental results presented in Fig. 5 assesses the importance of adding each potential function to the energy model defined in (3). This figure shows the mean Dice index results on the testing set of INbreast using the CRF and SSVM models. In particular, we show these results using several subsets of the potential functions "CNN", "DBN3 $\times$ 3", "DBN5 $\times$ 5", "GMM" , "Pairwise" and "Prior" presented in Sec. 3.3 (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). It is important to mention that the Dice index of our methodology using all potential functions on the training set of INbreast is 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 [28], falls to 0.85 using all potential functions for both models.

The comparison between the results from our methodology and other state-of-the-art results is shown in Tab. 1. This comparison is performed on the testing sets of DDSM-BCRP and INbreast, with the Dice index, average training time (for the whole training set) and testing time (per image), where our CRF and SSVM models have all potential functions: CNN+DBN3x3 + DBN5x5 + GMM + Prior + Pairwise. Notice that in this table, we only list the results available for the methods that use these publicly available databases because the great majority of papers published in this area have used subsets of the DDSM dataset and manual annotations that are not publicly available, which makes a direct comparison with these methods impossible.

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

Method	#Images	Dataset	Dice Index	Test Run.Time	Train Run. time
Proposed CRF model	116	INbreast	0.90	0.1s	360s
Proposed SSVM model	116	INbreast	0.90	0.8s	1800s
Cardoso et al. [36]	116	INbreast	0.88	?	?
Dhungel et al. [10]	116	INbreast	0.88	0.8s	?
Dhungel et al. [11]	116	INbreast	0.89	0.1s	?
Proposed CRF model	158	DDSM-BCRP	0.90	0.1s	383s
Proposed SSVM model	158	DDSM-BCRP	0.90	0.8s	2140s
Dhungel et al. [10]	158	DDSM-BCRP	0.87	0.8s	?
Dhungel et al. [11]	158	DDSM-BCRP	0.89	0.1s	?
Beller et al. [5]	158	DDSM-BCRP	0.70	?	?

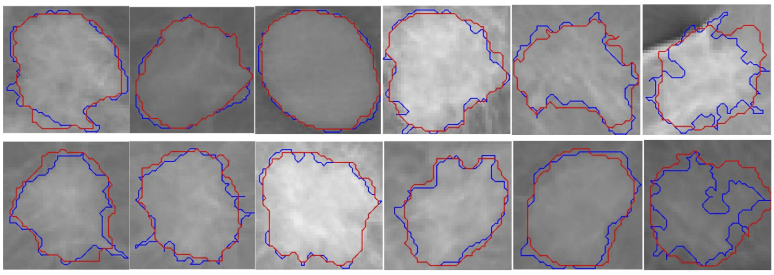


Fig. 6 Mass segmentation results produced by the CRF model on INbreast test images, where the blue curve denotes the manual annotation and red curve represents the automatic segmentation.

Finally, Fig. 6 shows examples of segmentation results produced by our CRF model on the test set of INbreast.

5 Discussion and Conclusions

The results from Fig. 5 explains the importance of each potential function used in the CRF and SSVM models, where it is clear that the CNN potential function provides the largest boost in performance. The addition of GMM and shape prior to deep learning models provides considerable improvements for both CRF and SSVM models. Another interesting observation is the fact that image preprocessing [28] appears to be important since it shows a substantial gain in terms of segmentation accuracy. The comparison with other methods in Table 1 shows that our methodology currently produces the best results for both databases, and the CRF and SSVM models hold comparable results in terms of segmentation accuracy. However, the comparison in terms of training and testing running times shows a significant advantage to the CRF model.

There are other important conclusions to make about the training and testing processes that are not displayed in these results: 1) we tried other types of CNN

structures, such as with different filter sizes, and we also tried to use more than one CNN model as additional potential functions, but the use of only one CNN with the structure detailed in Sec. 4.1 produced the best result in cross validation (the main issue affecting the CNN models is overfitting); 2) for the DBN models, we have also tried different input sizes (e.g., 7×7 patches), but the combinations of the ones detailed in Sec. 4.1 provided the best cross-validation results; and 3) the training for both the CRF and SSVM models estimates a much larger weight to the CNN potential function compared to other potential functions in Sec. 3.3, indicating that this is the most important potential function, but the CNN model alone (without CRF or SSVM) overfits the training data (with a Dice of 0.87 on test and 0.95 on training), so the structural prediction models serve as a regularizer to the CNN model. Finally, from the visual results in Fig. 6, we can see that our proposed CRF model produces quite accurate segmentation results when the mass does not show very sharp corners and cusps. We believe that the main issue affecting our method in these challenging cases is the limited size of the training sets in the DDSM-BCRP and INbreast datasets, which do not contain enough examples of such segmentations in order to allow an effective learning of a model that can deal with such complicated segmentation problems.

Acknowledgements

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).

References

1. N. Dhungel, G. Carneiro, and A. P. Bradley, "Deep learning and structured prediction for the segmentation of mass in mammograms," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pp. 605–612, Springer, 2015.
2. A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer statistics, 2008," *CA: a cancer journal for clinicians*, vol. 58, no. 2, pp. 71–96, 2008.
3. Y. Yuan, M. L. Giger, H. Li, K. Suzuki, and C. Sennett, "A dual-stage method for lesion segmentation on digital mammograms," *Medical physics*, vol. 34, p. 4180, 2007.
4. P. Rahmati, A. Adler, and G. Hamarneh, "Mammography segmentation with maximum likelihood active contours," *Medical image analysis*, vol. 16, no. 6, pp. 1167–1186, 2012.
5. M. Beller, R. Stotzka, T. O. Müller, and H. Gemmeke, "An example-based system to support the segmentation of stellate lesions," in *Bildverarbeitung für die Medizin 2005*, pp. 475–479, Springer, 2005.
6. J. G. Elmore, S. L. Jackson, L. Abraham, D. L. Miglioretti, P. A. Carney, B. M. Geller, B. C. Yankaskas, K. Kerlikowske, T. Onega, R. D. Rosenberg, *et al.*, "Variability in interpretive performance at screening mammography and radiologists characteristics associated with accuracy1," *Radiology*, vol. 253, no. 3, pp. 641–651, 2009.

7. I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.
8. M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th international workshop on digital mammography*, pp. 212–218, 2000.
9. J. J. Fenton, S. H. Taplin, P. A. Carney, L. Abraham, E. A. Sickles, C. D'Orsi, E. A. Berns, G. Cutter, R. E. Hendrick, W. E. Barlow, *et al.*, "Influence of computer-aided detection on performance of screening mammography," *New England Journal of Medicine*, vol. 356, no. 14, pp. 1399–1409, 2007.
10. N. Dhungel, G. Carneiro, and A. P. Bradley, "Deep structured learning for mass segmentation from mammograms," in *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 2950–2954, Sept 2015.
11. N. Dhungel, G. Carneiro, and A. P. Bradley, "Tree re-weighted belief propagation using deep learning potentials for mass segmentation from mammograms," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 760–763, April 2015.
12. G. Carneiro, B. Georgescu, S. Good, and D. Comaniciu, "Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 9, pp. 1342–1355, 2008.
13. J. Domke, "Learning graphical model parameters with approximate marginal inference," *arXiv preprint arXiv:1301.3193*, 2013.
14. I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, pp. 1453–1484, 2005.
15. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks.," in *NIPS*, vol. 1, p. 4, 2012.
16. Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.
17. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
18. C. Wang, N. Komodakis, and N. Paragios, "Markov random field modeling, inference and learning in computer vision and image understanding: A survey," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1610 – 1627, 2013.
19. M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching," in *Workshop on Artificial Intelligence and Statistics*, vol. 21, p. 97, Society for Artificial Intelligence and Statistics Np, 2003.
20. Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.
21. M. Szummer, P. Kohli, and D. Hoiem, "Learning crfs using graph cuts," in *Computer Vision–ECCV 2008*, pp. 582–595, Springer, 2008.
22. D. M. Catarious Jr, A. H. Baydush, and C. E. Floyd Jr, "Incorporation of an iterative, linear segmentation routine into a mammographic mass cad system," *Medical physics*, vol. 31, no. 6, pp. 1512–1520, 2004.
23. E. Song, L. Jiang, R. Jin, L. Zhang, Y. Yuan, and Q. Li, "Breast mass segmentation in mammography using plane fitting and dynamic programming," *Academic radiology*, vol. 16, no. 7, pp. 826–835, 2009.
24. S. Timp and N. Karssemeijer, "A new 2d segmentation method based on dynamic programming applied to computer aided detection in mammography," *Medical Physics*, vol. 31, no. 5, pp. 958–971, 2004.
25. A. Rojas Domínguez and A. K. Nandi, "Toward breast cancer diagnosis based on automated segmentation of masses in mammograms," *Pattern Recognition*, vol. 42, no. 6, pp. 1138–1148, 2009.

26. M. Yu, Q. Huang, R. Jin, E. Song, H. Liu, and C.-C. Hung, "A novel segmentation method for convex lesions based on dynamic programming with local intra-class variance," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 39–44, ACM, 2012.
27. S. Xu, H. Liu, and E. Song, "Marker-controlled watershed for lesion segmentation in mammograms," *Journal of digital imaging*, vol. 24, no. 5, pp. 754–763, 2011.
28. J. Ball and L. Bruce, "Digital mammographic computer aided diagnosis (cad) using adaptive level set segmentation," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 4973–4978, IEEE, 2007.
29. G. M. te Brake, N. Karssemeijer, and J. H. Hendriks, "An automatic method to discriminate malignant masses from normal tissue in digital mammograms," *Physics in medicine and biology*, vol. 45, no. 10, p. 2843, 2000.
30. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics*, vol. 28, no. 7, pp. 1455–1465, 2001.
31. J. A. Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*, vol. 3. Cambridge university press, 1999.
32. J. Shi, B. Sahiner, H.-P. Chan, J. Ge, L. Hadjiiski, M. A. Helvie, A. Nees, Y.-T. Wu, J. Wei, C. Zhou, *et al.*, "Characterization of mammographic masses based on level set segmentation with new image features and patient information," *Medical physics*, vol. 35, no. 1, pp. 280–290, 2007.
33. S. Nowozin and C. H. Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.
34. T. Meltzer, A. Globerson, and Y. Weiss, "Convergent message passing algorithms: a unifying view," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 393–401, AUAI Press, 2009.
35. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.
36. J. S. Cardoso, I. Domingues, and H. P. Oliveira, "Closed shortest path in the original coordinates with an application to breast cancer," *International Journal of Pattern Recognition and Artificial Intelligence*, 2014.