Hindawi Publishing Corporation
International Journal of Biomedical Imaging
Volume 2011, Article ID 846312, 10 pages
doi:10.1155/2011/846312

Research Article

Conditional Random Fields and Supervised Learning in Automated Skin Lesion Diagnosis

Paul Wighton,[1,2,3] Tim K. Lee,[1,2,3] Greg Mori,[1] Harvey Lui,[2,3] David I. McLean,[2] and M. Stella Atkins[1,2]

[1] *Department of Computing Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6*
[2] *Department of Dermatology and Skin Science, Photomedicine Institute, University of British Columbia and Vancouver Coastal Health Research Institute, Vancouver, BC, Canada V5Z 4E8*
[3] *Cancer Control Research Program and Integrative Oncology Department, BC Cancer Research Centre, Vancouver, BC, Canada V5Z 4E6*

Correspondence should be addressed to Paul Wighton, paul.wighton@gmail.com

Received 2 March 2011; Revised 21 May 2011; Accepted 23 July 2011

Academic Editor: Fei Wang

Many subproblems in automated skin lesion diagnosis (ASLD) can be unified under a single generalization of assigning a label, from an predefined set, to each pixel in an image. We first formalize this generalization and then present two probabilistic models capable of solving it. The first model is based on independent pixel labeling using maximum a-posteriori (MAP) estimation. The second model is based on conditional random fields (CRFs), where dependencies between pixels are defined using a graph structure. Furthermore, we demonstrate how supervised learning and an appropriate training set can be used to automatically determine all model parameters. We evaluate both models' ability to segment a challenging dataset consisting of 116 images and compare our results to 5 previously published methods.

1. Introduction

Incidence rates of melanoma are increasing rapidly in the western world, growing faster than any other cancer [1]. Since there is no effective therapy for patients with advanced melanoma [2], educational campaigns attempt to encourage high-risk individuals to undergo routine screening so that melanomas can be identified early while they are still easily treatable [3]. While worthwhile, these educational campaigns generate a large amount of referrals to dermatologists, whose services are already undersupplied [4].

Automated skin lesion diagnosis (ASLD) is expected to alleviate some of this burden. By acting as a screening tool, ASLD can reject obviously benign lesions, while referring more suspicious ones to an expert for further scrutiny. Most ASLD methods adopt the standard computer-aided diagnosis (CAD) pipeline illustrated in Figure 1. First an image is acquired with a digital dermoscope. Next, undesirable artifacts (such as hair or oil bubbles) are identified and, if necessary, replaced with an estimate of the underlying skin color. After this, the lesion is segmented, and discriminative features are then extracted. Finally, supervised learning is used to classify previously unseen images.

Our previous work demonstrated how the use of supervised learning, under the proper generalization (of assigning labels to pixels), was able to solve several tasks in this pipeline including detecting occluding hair, segmenting the lesions, and detecting the dermoscopic structure *pigment network* [5]. Our method was relatively simple; it labeled pixels in an image independently using modest features, linear discriminant analysis (LDA) for supervised dimensionality reduction, and maximum a-posteriori (MAP) estimation. Nevertheless, in spite of its simplicity, our model was able to perform comparably to other previously published, nongeneral methods for lesion segmentation [6–10] and hair detection [11].

In this paper, we seek to expand on this generalization by replacing the per-pixel (PP) estimation model with a conditional random field (CRF) model. The largest criticism levied at the PP approach is that pixels are labeled independently,

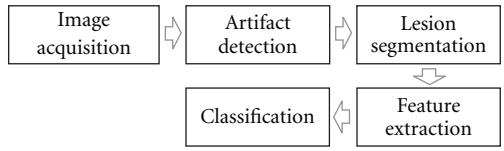

Figure 1: Typical computer aided diagnosis (CAD) pipeline usually adopted for automated skin lesion diagnosis (ASLD). Our goal is to (1) generalize the artifact detection, segmentation as well as a portion of the feature extraction stage into a single mathematical framework and (2) propose and evaluate probabilistic models which employ supervised learning to quickly and automatically "learn" to perform these tasks.

regardless of the label assigned to their neighbors. This assumption of independence is clearly not valid, as there is a high degree of correlation between neighboring pixels in any image (any image other than pure noise, i.e.). CRF-based models attempt to relax this assumption of independence by creating a graphical model which defines the dependencies between pixels.

In order to apply a CRF model, a parameter vector specifying the relative contribution of the input features is required. Often, these parameters are determined in an ad hoc fashion via trial and error. Since our goal is a *general* method, easily applicable to a variety of problems, it is crucial that these parameters be determined automatically based on observations. We, therefore, apply the maximum likelihood estimator for the parameter vector and describe a gradient-based method for its computation. We also address many practical considerations encountered during the implementation.

The paper is organized as follows: in Section 2, we briefly review previous work. In Section 3, we formulate the generalization in Section 3.1, review our previous PP model [5] in Section 3.2, and present our CRF model in Section 3.3. In Section 4, we present results. Finally, we conclude in Section 5.

2. Previous Work

Our original PP model was based on the work by Debeir et al. [12] who also attempts to generalize many tasks in ASLD. Our model was found to perform comparably to many published lesion segmentation algorithms including K-means++ (KPP) [6], J-image segmentation (JSEG) [7], dermatologist-like tumor area extraction algorithm (DTEA) [8], statistical region merging (SRM) [9], and threshold fusion (FSN) [10]. It also performed comparably to DullRazor [11] for detecting occluding hair and was able to identify the dermoscopic structure *pigment network*. Our PP model is briefly reviewed in Section 3.2; however, we refer readers to our previous study for further details, as well as a more comprehensive review of previous work in ASLD, including the methods against which we compare [5].

We defer the review of the CRF model until Section 3.3, where we examine it in detail.

3. Method

This section is divided into 3 parts. We begin in Section 3.1 by formalizing the generalization that is capable of performing a variety of tasks in ASLD. In Section 3.2, we briefly review our previous PP model [5]. Finally, in Section 3.3, we outline our CRF model.

3.1. The Generalization. We are given a set of observations $\{x^m, y^m\}$, consisting of images (x) and corresponding ground truths labeling (y). Using the notation of Szummer et al. [13], the superscript x^m or y^m indexes a specific image/labeling in the set and the subscript x_i or y_i indexes a specific pixel. Let N_I represent the number of images, and N_P^m represents the number of pixels in image x^m. An imageset can contain any number of channels (or features), which we denote by N_C. Valid values for each entry in the label field (y_i) are defined by the label set $L = \{l_1, \ldots, l_{N_L}\}$, where N_L is the number of possible labels.

Given our training set $\{x^m, y^m\}$, we use supervised learning to predict the label fields for previously unseen images.

Formally, we are given

$$\begin{aligned} &\{x^m, y^m\}; \quad m = 1, \ldots, N_I; \ \textit{i.i.d}, \\ &L = \{l_i\}; \quad i = 1, \ldots, N_L; \ l_i \in \mathbb{N}, \\ &x^m \in \mathbb{R}^{N_P^m \times N_c}, \\ &y^m \in L^{N_P^m}. \end{aligned} \tag{1}$$

And our task is to use the information in $\{x^m, y^m\}$ to infer the function $f : x \rightarrow y$ that produces the best possible label field.

3.2. The PP Model. In this section, we briefly review our per-pixel (PP) estimation model [5]. An overview of the training and testing phases of the model is illustrated in Figures 2 and 3, respectively. Under this model, we assign the most probable label to each pixel independently

$$y_i^* = \arg\max_{l_j} \left[P\left(y_i = l_j \mid x_i\right) \right]; \quad i = 1, \ldots, N_P, \tag{2}$$

Applying Bayes' rule and simplifying, we arrive at the standard maximum likelihood formulation

$$y_i^* = \arg\max_{l_j} \left[P\left(x_i \mid y_i = l_j\right) P(y_i) \right]; \quad i = 1, \ldots, N_P. \tag{3}$$

We model the posterior $P(x \mid y = l)$ as a set of N_L multivariate normal distributions $P(x \mid y = l_j) = N(\mu_{l_j}, \Sigma_{l_j})$, whose parameters $(\mu_{l_j}, \Sigma_{l_j})$ are estimated using the training set $\{x^m, y^m\}$. We model $P(y)$ as a discrete distribution. Let N_{Yi} represent the number of elements in y^m that assume the value l_i, then

$$P(y_i) = \frac{N_{Yi}}{\sum_{j=1}^{N_L} N_{Yj}}. \tag{4}$$

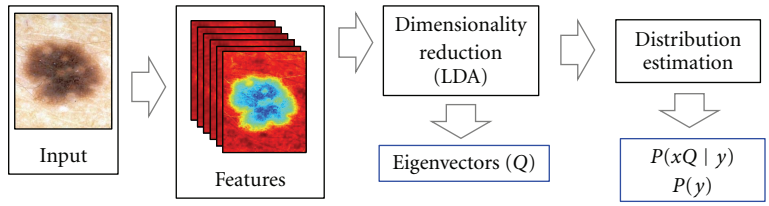

Figure 2: The training phase of our per-pixel (PP) model. Features are first computed, then the dimensionality of the featurespace is reduced using LDA. Posterior probabilities in this subspace are then estimated.

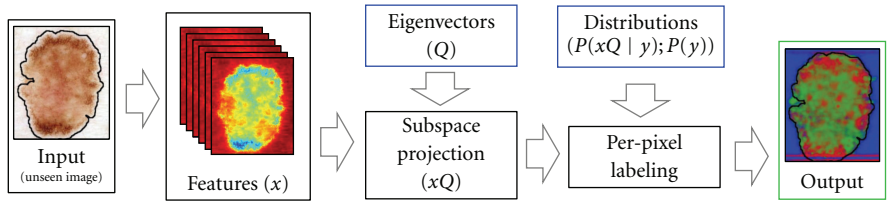

Figure 3: The testing phase of our per-pixel (PP) model. Features are computed as in the training phase. The projection Q is used to transform the features into the subspace determined in the training phase. Maximum a-posteriori estimation, using the posteriors estimated in the training phase, is then used to generate the label.

We also normalize the probabilities across the label set, which are later used as features in the CRF model. The normalized likelihood that a pixel i is associated with the label l_j is

$$\mathcal{L}_{i,j} = \frac{P\left(x_i \mid y_i = l_j\right)}{\sum_{k=1}^{N_L} P\left(x_i \mid y_i = l_k\right)}. \tag{5}$$

In order to examine the model's performance across the entire sensitivity/specificity range, we consider many thresholds T on $\mathcal{L}_{i,j}$ over the range $[0, 1]$ and label pixels accordingly.

As the number of channels (N_C) in the images grows, we perform supervised dimensionality reduction on the observations x to focus the predictive power of our dataset onto a smaller subset of parameters. Linear discriminant analysis (LDA) is used to determine the subspace of x which best separates the labels [14].

LDA performs an eigenvalue decomposition of a scatter matrix representing the ratio of between-class covariance to within-class covariance. It returns a matrix of eigenvectors $Q \in \mathbb{R}^{N_c \times N_L - 1}$ which projects observations (x) from N_C dimensions to $N_L - 1$

$$\begin{aligned} Q &= \operatorname{eig}\left(S_w^{-1} S_b\right), \\ S_w &= \sum_{i=1}^{N_L} \Sigma_{l_i}, \\ S_b &= \sum_{i=1}^{N_L} \left(\mu_{l_i} - \mu\right)\left(\mu_{l_i} - \mu\right)^T, \end{aligned} \tag{6}$$

where μ is the overall mean of x across all images and classes. Once the projection Q is determined, the posteriors are estimated, likelihoods are computed, and inference is performed in this subspace (xQ)

$$\begin{aligned} P\left(xQ \mid y = l_j\right) &= N\left(\mu_{l_j}^Q, \Sigma_{l_j}^Q\right); \quad j = 1, \ldots, N_L, \\ y_i^* &= \arg\max_{l_j}\left[P\left(x_i Q \mid y_i = l_j\right) P\left(y_i\right)\right]; \quad i = 1, \ldots, N_P, \end{aligned} \tag{7}$$

$$\mathcal{L}_{i,j} = \frac{P\left(x_i Q \mid y_i = l_j\right)}{\sum_{k=1}^{N_L} P\left(x_i Q \mid y_i = l_k\right)}, \tag{8}$$

where the superscript Q ($\mu_{l_j}^Q, \Sigma_{l_j}^Q$) is used to differentiate the label means/covariances in this subspace from the original space in which the observations were performed (μ_{l_j}, Σ_{l_j}).

3.3. The CRF Model. In this section, we seek to improve upon the PP model developed in previous work [5] and described in Section 3.2. We present an overview of conditional random fields (CRFs) in Section 3.3.1. In Section 3.3.2, we describe how the CRF parameters can be determined empirically using maximum likelihood estimation (MLE) [15]. In Section 3.3.3, we discuss practical considerations for finding these parameters, including how to estimate the partition function [16] and how to regularize the likelihood expression [15]. In Section 3.3.4, we solve the MLE formulation via gradient-based methods. An overview of the training and testing phases of our CRF model is illustrated in Figures 4 and 5, respectively.

3.3.1. Overview. The CRF model is an undirected graphical model that is naturally suited to represent and exploit the dependencies between observations, such as neighboring pixels in an image [15]. The probability that a label field y

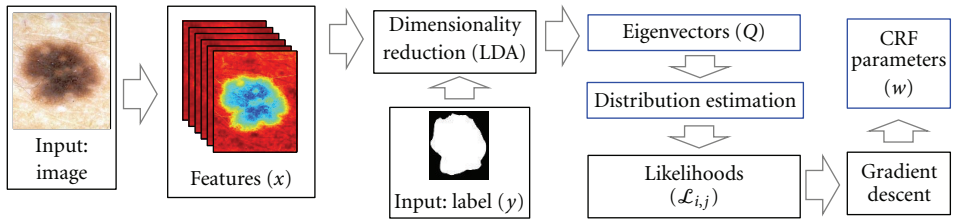

FIGURE 4: The training phase of our CRF model. We follow the same procedure as in our PP model up until the posteriors are estimated. We then calculate pixel likelihoods and use these as node features in our CRF model. We infer CRF parameters using gradient descent.

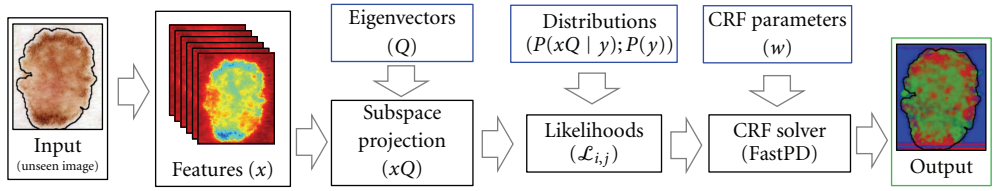

FIGURE 5: The testing phase of our CRF model. After the likelihoods are computed, we use the CRF parameters from the training phase, and the software FastPD to generate label fields.

is associated with the image x under model parameters w is given by

$$P(y \mid x; w) = \frac{1}{Z(x,w)} \exp(-E(y,x;w)), \tag{9}$$

where the function $Z(x,w)$, known as the partition function, is used to normalize the probabilities for given values of x and w

$$Z(x,w) = \sum_{y} \exp(-E(y,x;w)). \tag{10}$$

The energy function E represents the linear of combination of features employed by the model and is parameterized by the weight vector w

$$E(y,x;w) = \sum_{k=1}^{N_W} w_k \Phi_k(y,x). \tag{11}$$

Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the nodes (i.e., pixels) of an observation, $\mathcal{E}$ represents the *dependencies* between nodes (throughout this document, $\mathcal{E}$ is the 4-connected set of neighboring pixels), and the energy function E is the weighted sum of features $\Phi_i(y,x)$. Features can either operate over the nodes of the graph (Φ^V), or over its edges (Φ^E)

$$\begin{aligned} \Phi^V(y,x) &= \sum_{i \in \mathcal{V}} \phi(y_i, x_i), \\ \Phi^E(y,x) &= \sum_{(i,j) \in \mathcal{E}} \phi\left(y_i, y_j, x_i, x_j\right), \end{aligned} \tag{12}$$

In order for the model to be tractable, edge features $\Phi_i^E(y,x)$, and their corresponding weights must adhere to certain constraints. Let $\mathbb{E}$ represent the set of edge features. The following constraints must be satisfied [17]

$$\begin{aligned} w_i &> 0 \quad \forall i \in \mathbb{E}, \\ \phi^E\left(y_i, y_j, x_i, x_j\right) &= 0 \quad \forall \left(y_i, y_j\right) \quad \text{s.t. } y_i = y_j. \end{aligned} \tag{13}$$

Strictly speaking, the second constraint can be replaced with the more general constraint that edge feature functions be *submodular* [18]. However, throughout this document, we will impose this stricter constraint which can be interpreted as "an edge cost is only incurred across nodes with differing labels."

A CRF solver is one that, given observations x and parameters w, can find the most likely labeling y^*

$$y^* \longleftarrow \arg\max_{y} P(y \mid x; w). \tag{14}$$

We use the software FastPD [19, 20], which can exactly solve (14), under the constraints imposed above.

3.3.2. Determining MLE Parameters. Since the emphasis of our work is on a general model capable of performing a variety of tasks, it is crucial that model parameters (w) be determined automatically from training data via empirical means. In this section, we derive the partial derivatives of the likelihood function which can be used by gradient-based methods to compute w.

Since the observations $\{x^m, y^m\}$ are assumed to be independent. The likelihood of the data, given the set of parameters, is equal to the product of the probabilities in the observed set, under those parameters

$$\ell(w) = \prod_{m=1}^{N_I} P(y^m \mid x^m; w). \tag{15}$$

The maximum likelihood estimator is then

$$w^* = \arg\max_w \prod_{m=1}^{N_I} P(y^m \mid x^m; w). \tag{16}$$

If we can find the partial derivatives $\partial\ell/\partial w_i$, we can optimize w using gradient-based methods. We begin by expressing the likelihood function $\ell(w)$ in terms of w

$$\begin{aligned} w^* &= \arg\max_w \prod_{m=1}^{N_I} P(y^m \mid x^m; w) \\ &= \arg\max_w \sum_{m=1}^{N_I} \ln(P(y^m \mid x^m; w)) \\ &= \arg\max_w \sum_{m=1}^{N_I} (-E(y^m, x^m, w) - \ln(Z(x, w))) \\ &= \arg\min_w \sum_{m=1}^{N_I} \left(\sum_{k=1}^{N_W} w_k \Phi_k(y^m, x^m) \right. \\ &\qquad \left. + \ln\left[\sum_y \exp\left(-\sum_{k=1}^{N_W} w_k \Phi_k(y, x^m) \right) \right] \right). \end{aligned} \tag{17}$$

Solving for the partial derivatives, we get the following expression for the gradients of the likelihood function:

$$\frac{\partial \ell}{\partial w_i} = \sum_{m=1}^{N_I} \left(\Phi_i(y^m, x^m) + \frac{\sum_y -\Phi_i(y, x^m) \exp\left(-\sum_{k=1}^{N_w} w_k \Phi_k(y, x^m) \right)}{\sum_y \exp\left(-\sum_{k=1}^{N_w} w_k \Phi_k(y, x^m) \right)} \right). \tag{18}$$

However, we now come to an impasse. The second term of (18) would have us iterating over all possible label fields y. For a binary classification task over a modestly sized image of 256×128, this would require a summation over $2^{256\times128} \approx 2 \times 10^{9000}$ labelings. Clearly this is intractable, and we must resort to estimating this second term.

3.3.3. Practical Considerations. In order to derive CRF parameters with grid-structured models for even modestly sized images, a method to estimate the partition function is required. Inspired by [21], we employ one of the simplest estimation methods and approximate the partition function using saddle-point approximation (SPA) [16]

$$\begin{aligned} \sum_y \Phi(y, x) &\approx \Phi(y^*, x), \\ y^* &\longleftarrow \arg\max_y P(y \mid x; w). \end{aligned} \tag{19}$$

We also introduce an additional practical consideration. Since gradient-based methods will be used to determine w, we regularize the likelihood function ($\ell(w)$) by the squared L2 norm of the parameters [15] to penalize large weight vectors (since scalar multiples of a weight vector produce identical results). This makes the resulting likelihood function *strictly convex*. The regularized likelihood is then

$$\ell(w) = \sum_{m=1}^{N_I} \left(\sum_{k=1}^{N_W} w_k \Phi_k(y^m, x^m) - \ln \sum_y \exp \sum_{k=1}^{N_W} w_k \Phi_k(y, x^m) \right). - \frac{\|w\|^2}{2\sigma^2} \tag{20}$$

And the gradients become

$$\frac{\partial \ell}{\partial w_i} = \sum_{m=1}^{N_I} \left(\Phi_i(y^m, x^m) + \frac{\sum_y -\Phi_i(y, x^m) \exp\left(-\sum_{k=1}^{N_w} w_k \Phi_k(y, x^m) \right)}{\sum_y \exp\left(-\sum_{k=1}^{N_w} w_k \Phi_k(y, x^m) \right)} \right) - \frac{w_i}{\sigma^2}. \tag{21}$$

Which under SPA becomes

$$\frac{\partial \ell}{\partial w_i} \approx \sum_{m=1}^{N_I} (\Phi_i(y^m, x^m) - \Phi_i(y^*, x^m)) - \frac{w_i}{\sigma^2}. \tag{22}$$

3.3.4. Implementation. We are now ready to implement a gradient-based method to estimate the CRF parameter vector w. Given an initial weight vector w^0, the gradients of the likelihood function are estimated as per (22). These gradients are used to update the weight vector, which in turn is used to estimate a new set of gradients. This process is repeated until convergence.

We have observed (as does [21]) that gradient methods using saddle point approximation lead to oscillating behavior. Therefore, we keep a record of the best empirical set of parameters found, rather than the parameters of the final iteration. We also enforce the constraint from (13) that weights for edge-based features must remain positive.

In addition to the training set ($\{x, y\}$), the algorithm also requires an initial weight vector (w^0), a regularization factor (σ^2), a step size (γ), and termination conditions (convergence criteria: ϵ; maximum number of iterations: N_{itr}). The algorithm has been found to be robust to these additional parameters. Pseudocode of our implementation is presented in Algorithm 1.

4. Results

Previous work has demonstrated our model's ability to generalize to many applications [5]. Here, we focus on a single application (lesion segmentation) and present results for our two models. We also compare our results to 5 previously published methods (KPP [6], JSEG [7], DTEA [8], SRM [9], and FSN [10]).

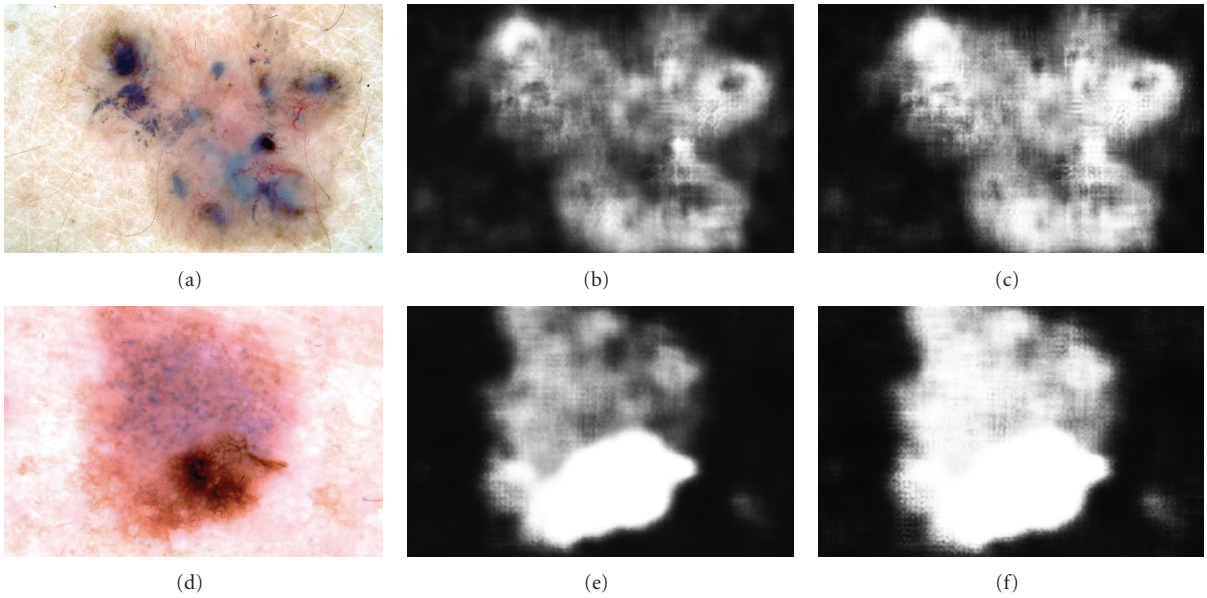

FIGURE 6: The effect of L* normalization on the segmentation likelihoods. left column: original dermoscopic image; middle: segmentation likelihoods ($\mathcal{L}_{i,\text{lesion}}$) before L* normalization; right: after L* normalization.

Require: $x, y, w^0, \sigma^2 > 0,\ \gamma < 0,\ \epsilon > 0,\ N_{itr} > 0$
  max $\leftarrow 0$
  for $i \leftarrow 1$ to N_{itr} **do**
    $g^i \leftarrow 0$
    $a \leftarrow 0$
    for $m \leftarrow 1$ to N_I **do**
      $y^* \leftarrow \arg\max_y P(y \mid x^m; w^{i-1})$
      $a \leftarrow a + \text{accuracy}(y^*, y^m)/N_I$
      $g^i \leftarrow g^i + \Phi(y^m, x^m) - \Phi(y^*, x^m) - w^{i-1}/(N_I\sigma^2)$
    end for
    if $a >$ max **then**
      max $\leftarrow a$
      $w^* \leftarrow w^{i-1}$
    end if
    $w^i \leftarrow w^{i-1} + \gamma g^i$
    for all $j \in \mathbb{E}$ **do**
      if $w_j^i < 0$ **then**
        $w_j^i \leftarrow 0$
      end if
    end for
    if $\|w^i - w^{i-1}\| < \epsilon$ **then**
      break
    end if
  end for
  return w^*

ALGORITHM 1: Calculating the CRF parameter vector w using gradient descent and saddle-point approximation.

The dataset consists of 116 images from dermoscopy atlases [22, 23], which were acquired by a several dermatologists in separate practices using differing equipment. The images have not been properly color calibrated. Since the goal was to create a difficult dataset, 100 of the 116 lesions were selected to be particularly challenging to segmentation algorithms [7]. We intentionally chose a simplistic featureset to emphasize the power of the models under consideration.

The features employed were 5 Gaussian, and 5 Laplacian of Gaussian filters applied a various scales ($\sigma = [1.25, 2.5, 5, 10, 20]$) in each channel of the image in CIE L*a*b* space. The responses of these filters represent the observations x (where $N_C = 30$). Each image was expertly segmented by a dermatologist. These ground truth labelings are denoted as y.

For all experiments, 10-fold cross-validation was employed. The dataset was randomly divided into 10 groups, and label fields for each group of images were determined using model parameters which were estimated from the observations in the 9 other groups. In both the PP and CRF models, all steps after the computation of features (refer to Figures 2 and 4) were included within the cross-validation loop including determining the projection Q, estimating the prior/posteriors, determining CRF parameter vector w, and so forth.

4.1. The PP Model. We begin by summarizing previous results on how our PP model faired on this dataset. A more detailed analysis, including the relative contribution of various aspects of the model (including features, dimensionality reduction, and classification method), can be found in our previous work [5].

Since that time, we have discovered that we can partially compensate for the lack of color calibration by subtracting the mean of the L* channel before computing features. While not as desirable as full color and lighting calibration [24], this procedure at least compensates for various camera exposure levels, as can be seen in the resulting PP likelihood maps in Figure 6 (as calculated by (8)). Figure 7 illustrates a ROC

Table 1: Comparison of our PP model's ability to segment lesions to our CRF model and 5 previously published methods.

		Performance			
		Method		PP (nearest pt.)	
Method	n	Sens	Spec	Sens	Spec
CRF	116	0.845	0.924	0.843	0.921
KPP [6]	116	0.765	0.770	0.941	0.763
JSEG [7]	91	0.627	0.987	0.677	0.980
DTEA [8]	116	0.597	0.986	0.638	0.985
SRM [9]	112	0.790	0.946	0.773	0.957
FSN [10]	116	0.808	0.934	0.814	0.939

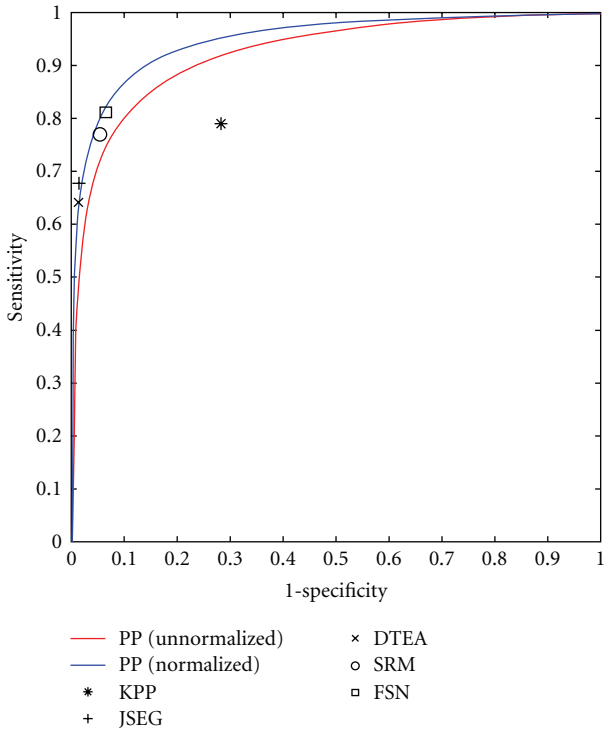

Figure 7: ROC curve comparing our PP model before normalization (red line) and after normalization (blue line) to 5 previously published methods.

curve comparing the performance of our PP model (before and after normalization) to the segmentation algorithms KPP [6], JSEG [7], DTEA [8], SRM [9], and FSN [10]. Our method performs comparably to JSEG, DTEA, SRM, and FSN and outperforms KPP although only KPP, DTEA, and FSN algorithms were able to generate results for all 116 images. Table 1 summarizes the results.

4.2. The CRF Model. As described in Section 3.3, the CRF model operates over an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and consists of node features ($\Phi^V(y,x)$) and edge features ($\Phi^E(y,x)$). The graph structure employed was the 4-connected set of neighboring pixels. Our featureset contains 2 features: one over the nodes and one over the edges. The node features are the likelihoods as computed by (8) of the PP model as in Section 4.1, and the edge features are set to the CIE L* intensity difference between neighboring pixels, if the labels of said pixels differ

$$\Phi_1^V(y,x) = \sum_{i\in\mathcal{V}} \frac{P(x_i Q \mid y_i)}{\sum_{j=1}^{N_L} P\left(x_i Q \mid y_i = l_j\right)},$$
$$\Phi_2^E(y,x) = \sum_{(i,j)\in\mathcal{E}} \left|L^*(x_i) - L^*\left(x_j\right)\right| \mathbf{1}_{y_i \neq y_j}, \tag{23}$$

where we use $\mathbf{1}_{y_i \neq y_j}$ to denote the indicator function (i.e., $\mathbf{1}_{y_i \neq y_j}$ evaluates to 1 if $y_i \neq y_j$; 0 otherwise)

While the method described in Section 3.3 is general enough to handle an arbitrary number of node and edge features, there are 2 reasons why we chose only one of each. To begin, we seek to make the comparison between the PP model and the CRF model as meaningful as possible. Using the likelihoods from the PP model as the node feature is an elegant way to evaluate the improvements realized by the CRF model. Note that with this choice of features, the CRF model with weight vector $w = [1, 0]$ gives identical results to the PP model. Additionally, the saddle-point method for approximating the partition function seems to degrade as the number of features increases. We note, however, that even in studies where the partition function can be computed exactly (because the CRF graph contains no loops), the loss incurred by such *piecewise training* methods is negligible [25].

Figure 8 compares some segmentations produced by the PP and CRF model. By relaxing the assumption of independence in the PP model, the CRF model is able to smooth over small areas of discontinuity, filling in "gaps" in segmentations, and removing noise. In Figures 8(a) and 8(c), the "holes" in the resulting PP segmentations do not manifest in the CRF segmentations (Figures 8(b) and 8(d)) due to the model's holistic search for the best label *field*, rather than best individual label. Additionally, while the PP model is already fairly robust to occluding hair (Figure 8(e)), the CRF model is even more robust, able to smooth over misclassifications due to artifacts.

We also tested the stability of the CRF model with respect to regularization and the hyperparameter σ^2. Varying σ^2 (to assume values in the range $[10^{-6}, \text{Inf}]$) had little effect on

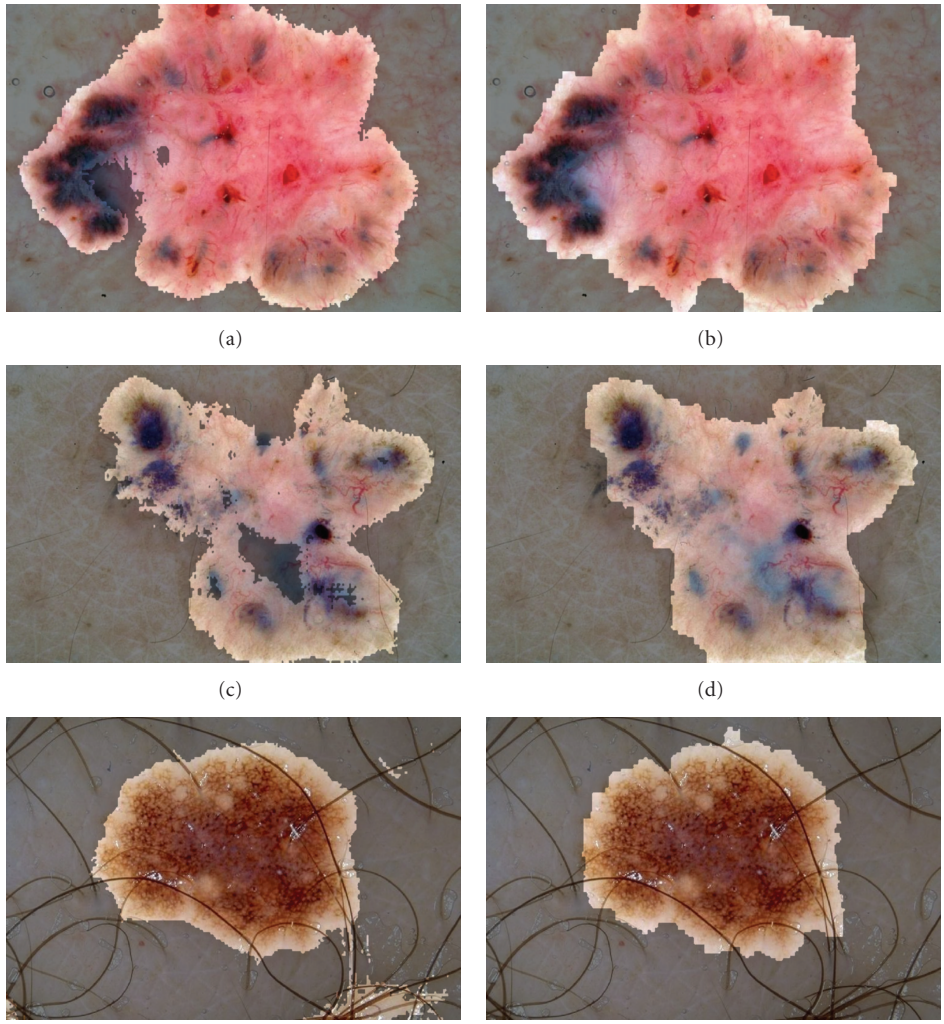

Figure 8: Comparing segmentations from our PP model (left) and CRF model (right). Since the CRF model relaxes the assumption of pixel independence in the PP model, it is able to smooth over local discontinuities. The result is better segmentations which fill in "holes" and remove "noise."

performance of the model on this particular dataset. In spite of the seemingly ineffectual nature of this parameter, we do not remove it from the model since the emphasis of this work is on *general* models for ASLD. The effect of σ^2 in general (over many tasks in ASLD) has yet to be determined.

While subjectively, the CRF model offers substantial improvements; objectively, the CRF model is a marginal improvement over the PP model. Figure 9 shows an ROC curve comparing the CRF's performance to that of the PP model and previously published methods, and Table 1 summarizes the results.

5. Conclusions

In this paper, we have generalized several common problems in ASLD into a single formulation. We also presented 2 probabilistic models capable of solving the formulation, and described how supervised learning can be used to determine all model parameters. Since the parameters for the resulting models can all be determined automatically from training data, it is hoped that these models can be applied quickly and effectively to a variety of relevant tasks in ASLD.

While both methods perform comparably to previously published methods, the qualitative improvements realized by CRF model aren't reflected in the quantitative score. Unlike the PP model, the CRF model does not assign labels to pixels independently. Rather, the CRF model selects the best label *field* to assign to an image. This allows the CRF model to fill in "holes" and smooth out noise that would otherwise appear.

The discrepancy between the objective and subjective performance of the CRF model implies that our evaluation metric (pixel-wise sensitivity and specificity) may be less than ideal. Therefore, future work will explore the use of alternate evaluation metrics [26, 27].

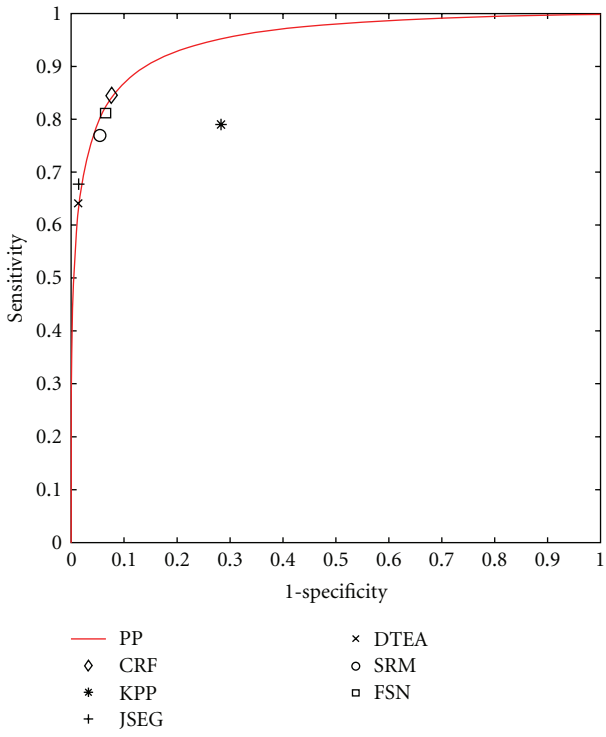

Figure 9: ROC curve comparing our CRF model (diamond) to our PP model (line) and 5 previously published methods.

Even though the models presented are competitive, there are many potential directions in which they can be improved upon even further. In our grid-structured CRF model, we must resort to approximating the partition function due to the computational complexity of calculating it exactly. Imposing a tree-based structure over the image [25] would enable the exact computation of the partition function via dynamic programming and should lead to more reliable CRF parameters. Replacing our gradient-based method for determining CRF parameters with a max-margin formulation [13] is another possible way to increase the reliability of the resulting parameters. We can also induce non-linearities into the model by replacing the linear dimensionality reduction step (LDA) with its nonlinear counterparts (i.e., KLDA [28]). Finally, the use of semi-supervised learning techniques may be used to decrease the cost of acquiring expertly annotated datasets [29].

Acknowledgments

This work was supported in part by the Canadian Institutes of Health Research (CIHR) Skin Research Training Centre, the Canadian Dermatology Foundation, and Collaborative Health Research Projects Program, a grant jointly funded by the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes of Health Research. The Authors would like to thank Dr. M. Emre Celebi for making available results of the alternative segmentation algorithms.

References

[1] C. Erickson and M. Driscoll, "Melanoma epidemic: facts and controversies," *Clinics in Dermatology*, vol. 28, no. 3, pp. 281–286, 2010.

[2] M. Lens and M. Dawes, "Global perspectives of contemporary epidemiological trends of cutaneous malignant melanoma," *British Journal of Dermatology*, vol. 150, no. 2, pp. 179–185, 2004.

[3] R. MacKie and D. Hole, "Audit of public education campaign to encourage earlier detection of malignant melanoma," *British Medical Journal*, vol. 304, no. 6833, pp. 1012–1015, 1992.

[4] J. Resneck and A. B. Kimball, "The dermatology workforce shortage," *Journal of the American Academy of Dermatology*, vol. 50, no. 1, pp. 50–54, 2004.

[5] P. Wighton, T. Lee, H. Lui, D. McLean, and M. Atkins, "Generalizing common tasks in automated skin lesion diagnosis," *IEEE Transactions on Information Technology in BioMedicine*, vol. 15, no. 4, pp. 622–629, 2011.

[6] H. Zhou, M. Chen, L. Zou et al., "Spatially constrained segmentation of dermoscopy images," in *Proceedings of the 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '08)*, pp. 800–803, Paris, France, 2008.

[7] M. E. Celebi, Y. Aslandogan, W. Stoecker, H. Iyatomi, H. Oka, and X. Chen, "Unsupervised border detection in dermoscopy images," *Skin Research and Technology*, vol. 13, no. 4, pp. 454–462, 2007.

[8] H. Iyatomi, H. Oka, M. E. Celebi et al., "An improved Internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm," *Computerized Medical Imaging and Graphics*, vol. 32, no. 7, pp. 566–579, 2008.

[9] M. E. Celebi, H. Kingravi, H. Iyatomi et al., "Border detection in dermoscopy images using statistical region merging," *Skin Research and Technology*, vol. 14, no. 3, pp. 347–353, 2008.

[10] M. E. Celebi, S. Hwang, H. Iyatomi, and G. Schaefer, "Robust border detection in dermoscopy images using threshold fusion," in *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP '10)*, pp. 2541–2544, Hong Kong, 2010.

[11] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, "Dullrazor®: a software approach to hair removal from images," *Computers in Biology and Medicine*, vol. 27, no. 6, pp. 533–543, 1997.

[12] O. Debeir, C. Decaestecker, J. Pasteels, I. Salmon, R. Kiss, and P. van Ham, "Computer-assisted analysis of epiluminescence microscopy images of pigmented skin lesions," *Cytometry*, vol. 37, no. 4, pp. 255–266, 1999.

[13] M. Szummer, P. Kohli, and D. Hoiem, "Learning CRFs using graph cuts," in *Proceedings of the 10th European Conference on Computer Vision (ECCV '08)*, pp. 582–595, Marseille, France, 2008.

[14] A. Martínez and A. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2002.

[15] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, p. 93, 2007.

[16] D. Geiger and F. Girosi, "Parallel and deterministic algorithms from MRFs: surface reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 5, pp. 401–412, 2002.

[17] H. Ishikawa, "Exact optimization for Markov random fields with convex priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1333–1336, 2003.

[18] V. Kolmogorov, "Primal-dual algorithm for convex Markov random fields," Microsoft Research MSR-TR-2005-117, 2005.

[19] N. Komodakis and G. Tziritas, "Approximate labeling via graph cuts based on linear programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1436–1453, 2007.

[20] N. Komodakis, G. Tziritas, and N. Paragios, "Performance vs computational efficiency for optimizing single and dynamic MRFs: setting the state of the art with primal-dual strategies," *Computer Vision and Image Understanding*, vol. 112, no. 1, pp. 14–29, 2008.

[21] S. Kumar, J. August, and M. Hebert, "Exploiting inference for approximate parameter learning in discriminative fields: an empirical study," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 153–168, Springer, New York, NY, USA, 2005.

[22] G. Argenziano and H. Soyer, *Interactive Atlas of Dermoscopy (Book and CD-ROM)*, Edra medical publishing and new media, Milan, Italy, 2000.

[23] H. Soyer and G. Argenziano, *Dermoscopy of Pigmented Skin Lesions. An Atlas based on the Consesnsus Net Meeting on Dermoscopy*, Edra medical publishing and new media, Milan, Italy, 2000.

[24] P. Wighton, T. Lee, H. Lui, D. McLean, and M. Atkins, "Chromatic aberration correction: an enhancement to the calibration of low-cost digital dermoscopes," *Skin Research and Technology*, vol. 17, no. 3, pp. 339–347, 2011.

[25] S. Nowozin, P. Gehler, and C. Lampert, "On parameter learning in CRF-based approaches to object class image segmentation," in *Proceedings of the 11th European Conference on Computer Vision (ECCV '10)*, pp. 98–111, Crete, Greece, 2010.

[26] M. Celebi, G. Schaefer, H. Iyatomi, W. Stoecker, J. Malters, and J. Grichnik, "An improved objective evaluation measure for border detection in dermoscopy images," *Skin Research and Technology*, vol. 15, no. 4, pp. 444–450, 2009.

[27] T. Lee, D. McLean, and M. Atkins, "Irregularity index: a new border irregularity measure for cutaneous melanocytic lesions," *Medical Image Analysis*, vol. 7, no. 1, pp. 47–64, 2003.

[28] D. Cai, X. He, and J. Han, "Efficient kernel discriminant analysis via spectral regression," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM '07)*, pp. 427–432, Omaha, Neb, USA, 2007.

[29] C. Lee, S. Wang, F. Jiao, D. Schuurmans, and R. Greiner, "Learning to model spatial dependency: semi-supervised discriminative random fields," *Advances in Neural Information Processing Systems*, vol. 19, pp. 793–800, 2007.