Approximate Parameter Learning in Conditional Random Fields: An Empirical Investigation[⋆]

Filip Korč and Wolfgang Förstner

University of Bonn, Department of Photogrammetry,
Nussallee 15, 53115 Bonn, Germany
filip.korc@uni-bonn.de,wf@ipb.uni-bonn.de
http://www.ipb.uni-bonn.de

Abstract. We investigate maximum likelihood parameter learning in Conditional Random Fields (CRF) and present an empirical study of pseudo-likelihood (PL) based approximations of the parameter likelihood gradient. We show, as opposed to [1][2], that these parameter learning methods can be improved and evaluate the resulting performance employing different inference techniques. We show that the approximation based on penalized pseudo-likelihood (PPL) in combination with the Maximum A Posteriori (MAP) inference yields results comparable to other state of the art approaches, while providing low complexity and advantages to formulating parameter learning as a convex optimization problem. Eventually, we demonstrate applicability on the task of detecting man-made structures in natural images.

Key words: Approximate parameter learning, pseudo-likelihood, Conditional Random Field, Markov Random Field

1 Introduction

Classification of image components in meaningful categories is a challenging task due to the ambiguities inherent into visual data. On the other hand, image data exhibit strong contextual dependencies in the form of spatial interactions among components. It has been shown that modeling these interactions is crucial to achieve good classification accuracy.

Conditional random field (CRF) provides a principled approach for combining local classifiers that allow the use of arbitrary overlapping features, with adaptive data-dependent label interaction. This formulation provides several advantages compared to the traditional Markov Random Field (MRF) model. Further, the restrictive assumption of conditional independence of data, made in the traditional MRFs, is relaxed in the CRF model.

CRFs [3] have been proposed in the context of segmentation and labeling of 1D text sequences. In [1], the concept of CRFs has been extended to graphs with

[⋆] Korč, F., Förstner, W. Approximate Parameter Learning in Conditional Random Fields: An Empirical Investigation, In *Pattern Recognition*, LNCS, Springer, 2008. Revised June 2008.

loops and made thus well applicable to problems in computer vision. The CRF that uses arbitrary discriminative classifiers to design the model potentials has been called the Discriminative Random Field (DRF).

In our previous work [4], we discussed the differences between a traditional MRF formulation and the DRF model, compared the performance of the two models and an independent sitewise classifier and demonstrated the application feasibility for the task of interpreting terrestrial images of urban scenes. Further, we presented preliminary results suggesting the potential for performance improvement.

Exact computation of the likelihood of CRF model parameters is in general infeasible for graphs with grid or irregular topology. For this reason, developing effective parameter learning methods is the crucial part in applying CRFs in computer vision. In [5], a model modification is described resulting in parameter learning formulated as a convex optimization problem. Learning/inference coupling is studied in [2]. Learning in CRFs can be accelerated using Stochastic Gradient Methods [6] and Piecewise Pseudo-likelihood [7]. A semi-supervised learning approach to learning in CRF can be found in [8]. In this work, we empirically investigate approximate parameter learning methods based on pseudo-likelihood (PL). We show that these methods yield results comparable to other state of the art approaches to parameter learning in CRFs, while providing desirable convergence behavior independent of the initialization

2 Conditional Random Field

CRFs are used in a *discriminative* framework to model the posterior over the labels given the data. In other words, let $\mathbf{y}$ denote the observed data from an input image, where $\mathbf{y} = \{\mathbf{y}_i\}_{i \in S}$, $\mathbf{y}_i$ is the data from the ith site, and S is the set of sites. Let the corresponding labels at the image sites be given by $\mathbf{x} = \{x_i\}_{i \in S}$. We review the CRF formulation in the context of binary classification on 2D image lattices. A general formulation on arbitrary graphs with multiple class labels is described in [9]. Thus, we now have $x_i \in \{-1, 1\}$ for a binary classification problem. In the considered CRF framework, the posterior over the labels given the data is expressed as,

$$P(\mathbf{x}|\mathbf{y}) = \frac{1}{Z} \exp\left(\sum_{i \in S} A_i(x_i, \mathbf{y}) + \sum_{i \in S} \sum_{j \in N_i} I_{ij}(x_i, x_j, \mathbf{y}) \right) \tag{1}$$

The CRF model in Eq. (1) captures the class association A_i at individual sites i with the interactions I_{ij} in the neighboring sites ij. The parameter dependent term $Z(\boldsymbol{\theta})$ is the normalization constant (or partition function) and is in general intractable to compute. N_i is the set of neighbors of the image site i.

Both, the unary association potential A_i and the pairwise interaction potential I_{ij} can be modeled as arbitrary unary and pairwise classifiers [5]. In this paper, as in [5][2] and our previous work [4], we use a logistic function $\sigma(t) = 1/(1 + e^{-t})$ to specify the local class posterior, i.e., $A_i(x_i, \mathbf{y}) = \log P'(x_i|\mathbf{y}) =$

$\log\sigma(x_i\mathbf{w}^T\boldsymbol{h}_i(\mathbf{y}))$ where the parameters $\mathbf{w}$ specify the classifier for individual sites. Here, $\boldsymbol{h}_i(\mathbf{y})$ is a sitewise feature vector, which has to be chosen such that a high positive weighted sum $\mathbf{w}^T\boldsymbol{h}_i(\mathbf{y})$ supports class $x_i = 1$. Similarly, to model I_{ij} we use a pairwise classifier of the following form: $I_{ij}(x_i, x_j, \mathbf{y}) = x_i x_j \mathbf{v}^T\boldsymbol{\mu}_{ij}(\mathbf{y})$. Here, the parameters $\mathbf{v}$ specify the classifier for site neighborhoods. $\boldsymbol{\mu}_{ij}(\mathbf{y})$ is a feature vector similarly being able to support or suppress the identity $x_i x_j = 1$ of neighboring class labels. We denote the unknown CRF model parameters by $\boldsymbol{\theta} = \{\mathbf{w}, \mathbf{v}\}$. In the following we assume the random field in Eq. (1) to be homogeneous and isotropic. Hence we drop the subscripts and use the notation A and I.

3 Parameter Learning

We learn the parameters $\boldsymbol{\theta}$ of the CRF model in a supervised manner. Hence, we use training images and the corresponding ground-truth labeling. We use standard maximum likelihood approach and, in principle, maximize the conditional likelihood $P(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ of the CRF model parameters. However, this would involve the evaluation of the partition function Z which is in general NP-hard. To overcome the problem, we may either use sampling techniques or approximate the partition function. As in [5], we use the pseudo-likelihood (PL) approximation $P(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}) \approx \prod_{i\in S} P(x_i|\mathbf{x}_{N_i},\mathbf{y},\boldsymbol{\theta})$ [10],[11], which is characterized by its relatively low computational complexity.

It has been observed [5] that this approximation tends to overestimate the interaction parameters, causing the MAP estimate of the field to be a poor solution. To overcome the difficulty, they propose to adopt the Bayesian viewpoint and find the maximum a posteriori estimate of the parameters by assuming a Gaussian prior over the parameters such that $P(\boldsymbol{\theta}|\tau) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0},\tau^2\mathbf{I})$ where $\mathbf{I}$ is the identity matrix. Thus, given M independent training images, we determine $\boldsymbol{\theta}$ from

$$\hat{\boldsymbol{\theta}}^{ML} \approx \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{m=1}^{M}\prod_{i\in S} P(x_i^m|\mathbf{x}_{N_i}^m,\mathbf{y}^m,\boldsymbol{\theta})P(\boldsymbol{\theta}|\tau)$$

or equivalently from the log likelihood

$$\hat{\boldsymbol{\theta}}^{ML} \approx \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{m=1}^{M}\sum_{i\in S}\left(A(x_i,\mathbf{y},\mathbf{w}) + \sum_{j\in N_i} I(x_i,x_j,\mathbf{y},\mathbf{v}) - \log z_i\right) - \frac{1}{2\tau^2}\mathbf{v}^T\mathbf{v} \tag{2}$$

where

$$z_i = \sum_{x_i\in\{-1,1\}} \exp\{A(x_i,\mathbf{y},\mathbf{w}) + \sum_{j\in N_i} I(x_i,x_j,\mathbf{y},\mathbf{v})\}$$

As stated in [5], if τ is given, the problem in Eq. (2) is convex with respect to the model parameters and can be maximized using gradient ascent. We note that it is an approximation of the true likelihood gradient that is now being computed. We implement a gradient ascent method variation with exact line search and maximize for different values of τ.

In our experiments, we adopt two methods of parameter learning. In the first set of experiments, we learn the parameters of the CRF using a uniform prior over the parameters in Eq. (2), i.e., $\tau = \infty$. This approach is referred to as the pseudo-likelihood (PL) learning method. Learning technique in the second set of experiments, where Gaussian prior over the CRF model parameters is used, is denoted as the penalized pseudo-likelihood (PPL) learning method. We specify further details of the PPL learning together with the experiments.

Discrete approximations of the partition function based on the Saddle Point Approximation (SPA) [12], Pseudo-Marginal Approximation (PMA) [13] and the Maximum Marginal Approximation (MMA) are described in [2]. A Markov Chain Monte Carlo (MCMC) sampling inspired method proposed in [14] and called the Contrastive Divergence (CD), is another way to deal with the combinatorial size of the label space. In our experiments we compare the PL based methods with these approaches to parameter learning.

4 Experiments

To analyze the learning and inference techniques described in the previous section, we applied the CRF model to a binary image restoration task. The aim of these experiments is to recover correct labeling from corrupted binary images. We use the data that has been used in learning and inference experiments in [5],[2]and compare our results with those published in the above mentioned works.

Four base images, see the bottom row in Fig. 1, 64×64 pixels each are used in the experiments. Two different noise models are employed: Gaussian noise and class dependent bimodal (two mixtures of two Gaussians) noise. Details of the noise model parameters are given in [5]. For each noise model, 10 out of 50 noisy images from the left most base image in Fig. 1 are used as the training set for parameter learning. The rest 190 noisy images are used for testing.

The unary and pairwise features are defined as: $\mathbf{h}_i(\mathbf{y}) = [1, I_i]^T$ and $\boldsymbol{\mu}_{ij}(\mathbf{y}) = [1, |I_i - I_j|]^T$ respectively, where I_i and I_j are the pixel intensities at the site i and the site j. Hence, the parameter $\mathbf{w}$ and $\mathbf{v}$ are both two-element vectors, i.e., $\mathbf{w} = [w_0, w_1]^T$, and $\mathbf{v} = [v_0, v_1]^T$.

4.1 Optimization

Finding optimal parameters of the CRF model means solving convex optimization problem in Eq. (2). For this purpose, we implement a variation of the gradient ascent algorithm with exact line search. For the computation of the numerical gradient we use the spacing 0.001 between points in each direction.

Plots in Fig. 2ac show negative logarithm of the objective function of the optimization problem in Eq. (2). Results in Fig. 2ab correspond to the objective function, where no prior over the parameters is used, i.e., to the PL learning

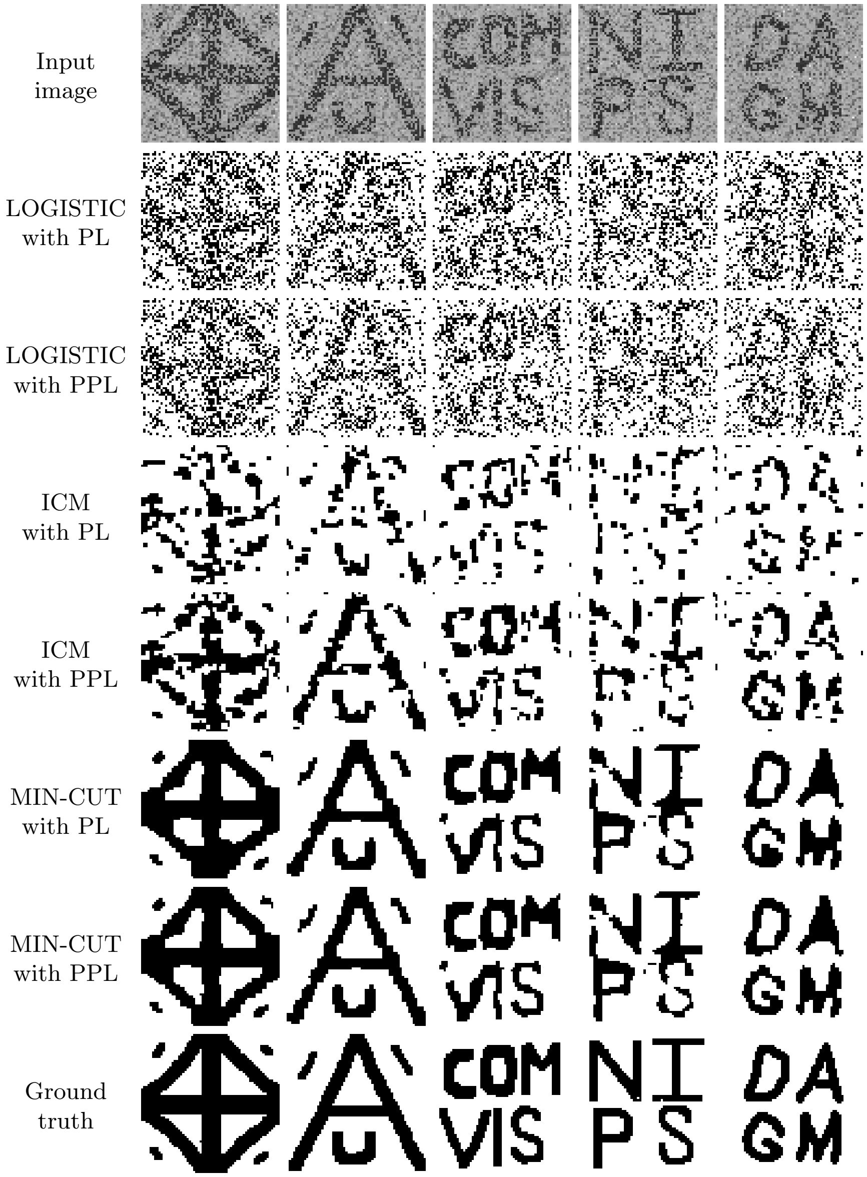
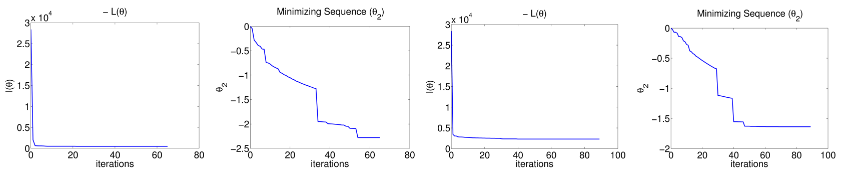


Fig. 1. Image restoration results for synthetic images corrupted with bimodal noise. Results for different combinations of parameter learning (PL: Pseudo-likelihood, PPL: Penalized Pseudo-likelihood) and inference methods (MIN-CUT: min-cut/max-flow algorithm, ICM: Iterated Conditional Modes, LOGISTIC: logistic classifier) are shown. Training and test data courtesy Sanjiv Kumar (4 image columns on the left).



(a) (b) (c) (d)

Fig. 2. Parameter learning using gradient ascent with exact line search. Plots of the negative logarithm of the approximated likelihood of the model parameters (a,c) and plots of the model parameter updates (b,d) for the PL parameter learning (a,b) and the PPL parameter learning (c,d).

method. Results in Fig. 2cd show learning where parameters are penalized by imposing a Gaussian prior. This corresponds to the PPL learning method.

Fig. 2bd shows minimizing sequences for the model parameter w_1. On this example, we illustrate that the model parameters change their values significantly although the criterion value does not decrease much compared to the initial iterations. This observation motivates the employment of *exact* optimization. An inexact approach, commonly used in practice, where the step length is chosen to approximately minimize the criterion along the chosen ray direction, stops the computation far from optimum in this case.

4.2 PL and PPL Parameter Learning

In the following, we first adopt the PPL learning approach and investigate different combinations of model parameter priors and values of the parameter τ. Gaussian priors over the following four combinations of parameters: $\{\mathbf{w}\}$, $\{\mathbf{v}\}$, $\{v1\ w1\}$, $\{\boldsymbol{\theta}\}$ are used in our experiments, where in each case uniform prior is used for the rest of the parameters. Further, we run the PPL parameter learning for the following values of the prior parameter $\tau = \{1, 0.1, 0.01, 0.001, 0.0001\}$.

We used 10 training images corrupted with both the Gaussian and the bimodal noise to learn the model parameters and, in this case, evaluated the method on all 200 images. Tab. 1 summarizes the experiment for the bimodal noise model by showing the resulting pixelwise classification errors obtained by the min-cut/max-flow algorithm (MIN-CUT) [15][16].

In accordance with [5], PPL learning with prior over the interaction parameters $\mathbf{v}$ together with the MIN-CUT inference yields the lowest classification error for both noise models. In addition to this, for the Gaussian noise model we find that also learning with prior over all the parameters $\boldsymbol{\theta}$ yields comparable classification error. As opposed to [5], $\tau = 0.1$ yields the lowest classification errors in case of the bimodal noise. Further, as opposed to [5], we find that prior parameter value $\tau = 1$ with the Gaussian noise model yields the best results.

We now employ the parameter prior and the value of τ identified in the previous experiment and validate our results by learning on 10 images randomly selected from the training set and subsequently testing on 190 images from the

Table 1. Pixelwise classification errors (%) on 200 images. Columns show combinations of model parameter priors and rows values of the prior parameter τ. See text for more.

Parameter	Parameter Prior			
τ	$\{\mathbf{w}\}$	$\{\mathbf{v}\}$	$\{v1\ w1\}$	$\{\boldsymbol{\theta}\}$
1	11.27	11.03	11.25	11.20
0.1	21.96	7.42	21.60	19.65
0.01	26.09	**6.25**	22.15	22.15
0.001	23.04	16.55	22.15	22.15

Table 2. Pixelwise classification errors (%) on 190 test images. Columns show parameter learning methods used with two noise models. KH'06 stands for the results published in [5]. Mean ± standard deviation over 10 experiments is given for our results.

	Learning Method			
	Gaussian Noise		Bimodal Noise	
	PL	PPL	PL	PPL
KH'06	3.82	2.30	17.69	6.21
ours	**2.55** ± 0.02	2.54 ± 0.02	**5.68** ± 0.05	**5.64** ± 0.04

test set. For every scenario we run the experiment 10 times and eventually give the mean together with the standard deviation. Tab. 2 summarizes the experiment and illustrates that our PPL learning improves results reported in [5].

We now adopt the PL parameter learning method and evaluate the approach in combination with the MIN-CUT inference. As illustrated in Tab. 2, we improve the performance of PL learning for both the Gaussian and the bimodal noise model. We attribute this improvement to the employment of exact approach to the optimization.

In our experiments, we compute a numerical gradient of the approximated likelihood. The time needed for learning could be decreased by computing exact gradient of the approximated likelihood. Employing an inexact line search would accelerate learning at the cost of approximate solution. We maintain that the learning time is in this case to a great extent initialization dependent.

Table 3. Pixelwise classification errors (%) on 190 test images. Rows show inference techniques and columns show parameter learning methods used with two noise models. Mean ± standard deviation over 10 experiments is given. MIN-CUT inference is implemented in C, the rest of the algorithms are implemented in Matlab.

	Learning Method				Inference
Inference	Gaussian Noise		Bimodal Noise		Time
Method	PL	PPL	PL	PPL	(sec)
LOGISTIC	15.30 ± 0.06	15.30 ± 0.06	30.52 ± 0.26	28.44 ± 0.01	0.002
ICM	4.33 ± 0.01	3.72 ± 0.01	22.52 ± 0.07	13.66 ± 0.15	0.100
MIN-CUT	**2.55** ± 0.02	**2.54** ± 0.02	**5.68** ± 0.05	**5.64** ± 0.04	0.018
Learning Time (sec)	42 ± 7	45 ± 8	25 ± 4	29 ± 4	

At last, we compare results of a logistic classifier (LOGISTIC), Iterated Conditional Modes (ICM) [17] and the MIN-CUT inference for the case of parameters learned through both the PL and the PPL method and for the both noise models. In our experiments, MIN-CUT inference yields the lowest classification error for both learning approaches. The experiment is summarized in Tab. 3 and typical classification results are further illustrated in Fig. 1.

4.3 Comparison of Learning Methods

For the MAP MIN-CUT inference, we compare our parameter learning with other state of the art learning methods proposed in [2] and mentioned in Sec. 3. We summarize the comparison in Tab. 4.

Table 4. Pixelwise classification errors (%) on 190 test images. Rows show parameter learning procedures and columns show two different noise models. KH'05 stands for the results published in [2]. Mean ± standard deviation over 10 experiments is given for our results.

	Gaussian noise MIN-CUT	Bimodal noise MIN-CUT	Learning time (Sec)
MMA, KH'05	34.34	26.53	636
PL, KH'05	3.82	17.69	300
CD, KH'05	3.78	8.88	207
PMA, KH'05	2.73	6.45	1183
SPA, KH'05	2.49	5.82	82
PL, ours	2.55 ± 0.02	**5.68** ± 0.05	**42** ± 7
PPL, ours	2.54 ± 0.02	**5.64** ± 0.04	**45** ± 8

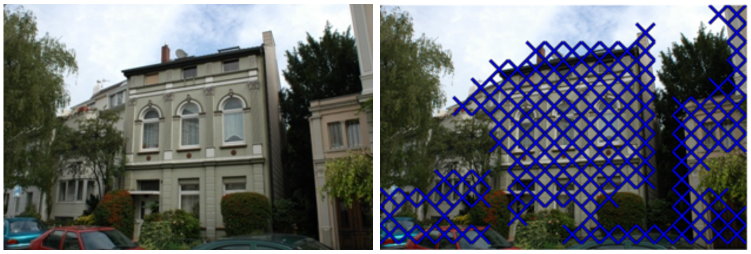
It was found in [2] that for MAP inference SPA based learning is the most accurate as well as time efficient. However, it was also showed that this approximation leads to a limit cycle convergence behavior dependent on the parameter initialization. As the convergence is not guaranteed, a parameter selection heuristics has to be chosen for the oscillatory case. This is the main drawback of the approximation.

In Tab. 4, we show that MAP inference with PPL based learning yields slightly better results compared to SPA learning while providing low complexity and advantages to formulating parameter learning as a convex optimization problem. In this case, the problem can be solved, very reliably and efficiently, drawing upon the benefits of readily available methods for convex optimization.

4.4 Natural Images

We demonstrate applicability on the task of detection of man-made structures in natural images and show preliminary results on real data. Our intention in this experiment is to label each site of a test image as *structured* or *non-structured*.

We divide our test images, each of size 3008 × 2000 pixels, into non-overlapping blocks, each of size 150 × 150 pixels, that we call image sites. For each image site i, a 1-dimensional single-site feature is computed as a linear combination of gradient magnitude and orientation based features. In the current setup, we reduce the CRF parameter learning to the determination of the interaction parameter w_0, where the rest of the parameters is fixed. We choose the values of the parameters that by observation yield the best performance on a test set of 15 images. See Fig. 3 for illustration.



(a) Input Image (b) CRF

Fig. 3. (a) Input Image. (b) Man-made structure detection result using the CRF model. Man-made structure is denoted by blue crosses superimposed on the original image.

5 Conclusion

We investigate maximum likelihood parameter learning in Conditional Random Fields (CRF) and present an empirical study of pseudo-likelihood (PL) based approximations of the parameter likelihood gradient. We show that the approximation based on penalized pseudo-likelihood (PPL) in combination with the Maximum A Posteriori (MAP) inference yields state of the art performance, while providing low complexity and desirable, initialization independent convergence. Eventually, we demonstrate the applicability of the method to the task of detecting man-made structures in natural images.

We are currently exploring further ways of efficient parameter learning in the CRFs on grid graphs and on graphs with general neighborhood system.

Acknowledgments. The authors would like to thank to V. Kolmogorov for the min-cut code and S. Kumar for the training and test data. The first author was supported by the EC Project FP6-IST-027113 eTRIMS.

References

1. Sanjiv Kumar and Martial Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proc. of the 9th IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 2003.

2. Sanjiv Kumar, Jonas August, and Martial Hebert. Exploiting inference for approximate parameter learning in discriminative fields: An empirical study. In *5th International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2005.
3. John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289, 2001.
4. Filip Korč and Wolfgang Förstner. Interpreting terrestrial images of urban scenes using Discriminative Random Fields. In *Proc. of the 21st Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2008.
5. Sanjiv Kumar and Martial Hebert. Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–201, June 2006.
6. S.V. N. Vishwanathan, Nicol N. Schraudolph, Mark W. Schmidt, and Kevin Murphy. Accelerated training of conditional random fields with stochastic gradient methods. In William W. Cohen and Andrew Moore, editors, *Proc. of the 24th International Conf. on Machine Learning*, volume 148, pages 969–976. ACM Press, 2006.
7. Charles Sutton and Andrew McCallum. Piecewise pseudolikelihood for efficient training of conditional random fields. In Zoubin Ghahramani, editor, *International Conference on Machine learning (ICML)*, volume 227, pages 863–870, 2007.
8. Chi-Hoon Lee, Shaojun Wang, Feng Jiao, Dale Schuurmans, and Russell Greiner. Learning to model spatial dependency: Semi-supervised discriminative random fields. In Bernhard Schölkopf, John Platt, and Thomas Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 793–800. MIT Press, 2007.
9. Sanjiv Kumar and Martial Hebert. Multiclass discriminative fields for parts-based object detection. In *Snowbird Learning Workshop*, 2004.
10. Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, September 1975.
11. Julian Besag. Efficiency of pseudo-likelihood estimation for simple gaussian fields. *Biometrika*, 64:616–618, 1977.
12. D. Geiger and F. Girosi. Parallel and deterministic algorithms from mrfs: Surface reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(5):401–412, 1991.
13. Andrew McCallum, Khashayar Rohanimanesh, and Charles Sutton. Dynamic conditional random fields for jointly labeling multiple sequences. In *NIPS Workshop on Syntax, Semantics, and Statistics*, December 2003.
14. Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.
15. D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society*, 51(2):271–279, 1989.
16. Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.
17. Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.