Appearing in *IEEE International Workshop on Machine Learning for Signal Processing*, São Luís, Brazil, Sep. 2004.

# SIGN DETECTION IN NATURAL IMAGES WITH CONDITIONAL RANDOM FIELDS

Jerod Weinman, Allen Hanson, and Andrew McCallum
{weinman,hanson,mccallum}@cs.umass.edu
Department of Computer Science
University of Massachusetts-Amherst

**Abstract. Traditional generative Markov random fields for segmenting images model the image data and corresponding labels jointly, which requires extensive independence assumptions for tractability. We present the conditional random field for an application in sign detection, using typical scale and orientation selective texture filters and a nonlinear texture operator based on the grating cell. The resulting model captures dependencies between neighboring image region labels in a data-dependent way that escapes the difficult problem of modeling image formation, instead focusing effort and computation on the labeling task. We compare the results of training the model with pseudo-likelihood against an approximation of the full likelihood with the iterative tree reparameterization algorithm and demonstrate improvement over previous methods.**

## INTRODUCTION

Image segmentation and region labeling are common problems in computer vision. In this work, we seek to identify signs in natural images by classifying regions according to their textural properties. Our goal is to integrate with a wearable system that will recognize any detected signs as a navigational aid to the visually impaired. Generic sign detection is a difficult problem. Signs may be located anywhere in an image, exhibit a wide range of sizes, and contain an extraordinarily broad set of fonts, colors, arrangements, etc. For these reasons, we treat signs as a general texture class and seek to discriminate such a class from the many others present in natural images.

The value of context in computer vision tasks has been studied in various ways for many years. Two types of context are important for this problem: label context and data context. In the absence of label context, local regions are classified independently, which is a common approach to object detection. Such disregard for the (unknown) labels of neighboring regions often leads to isolated false positives and missing false negatives. The absence of data context means ignoring potentially helpful image data from any neighbors of the region being classified. Both contexts are simultaneously important. For instance, since neighboring regions often have the same label, we could penalize label discontinuity in an image. If such regularity is imposed without regard for the actual data in a region and local evidence for a label is weak, then continuity constraints would typically override the local data. Conversely, local region evidence for a "sign" label could be weak, but a strong edge in the adjoining region might bolster belief in the presence of a sign at the site because the edge indicates a transition. Thus, considering both the labels *and* data of neighboring regions is important for predicting labels. This is exactly what the conditional random field (CRF) model provides.

The advantage of the discriminative contextual model over a generative one for detection tasks has recently been shown in [8]. We demonstrate a training method that improves prediction results, and we apply the model to a challenging real-world task. First the details of the model and how it differs from the typical random field are described, followed by a description of the image features we use. We close with experiments and conclusions.

## RANDOM FIELDS

### Model

For many computer vision tasks, the prior probability of the data being observed is inconsequential. Images happen. We are primarily interested in what may be inferred when *given* the images. However, probability distributions over labels $\mathbf{y}$ and an image $\mathbf{x}$ have traditionally been modeled jointly, with the image prior probability being ignored at classification time. For that reason, generative joint models require unnecessary modeling effort and more computation than their conditional counterparts.

Markov random fields are probability distributions parameterized by a graph topology $G = (V, E)$. For tractability reasons, typical generative random fields treat the interaction between local data and its label independently of the interaction between neighboring labels. The joint distribution is thus factored into the prior on label assignments and the probability of locally observed data, conditioned on the single site label:

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{x} \mid \mathbf{y})\, p(\mathbf{y}) \triangleq \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{y}_C) \prod_{v \in V} \psi_v(y_v, x_v), \tag{1}$$

where $\psi(\cdot)$ are compatibility functions, $Z$ is a normalizing constant making the expression a probability distribution, $\mathcal{C}$ is a family of cliques of the graph, and $\mathbf{y}_C$ are the variables in a given clique $C \subset V$. In this model, objects $x$ (e.g., patch statistics, salient features, etc.) from each class $y \in \mathcal{Y}$ are generated by a class-conditional probability distribution $p(x \mid y)$. This requires not only a model for every class we wish to distinguish, but an accurate generative background model even for classes of no interest; a non-trivial task because the real world contains a myriad of image "classes" (region types, textures,

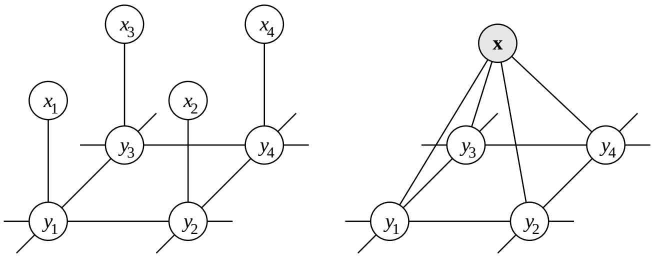


Figure 1: LEFT: Traditional joint random field over data $\mathbf{x}$ and labels $\mathbf{y}$ (cf. Eq. 1). RIGHT: Conditional random field where data is observed (cf. Eq. 2).

objects, etc.). In short, it is generally more difficult to explain the processes that generate class data than it is to model the boundaries between classes. In the latter approach, only boundaries among classes of interest must be distinguished, with the remainder easily collapsing into a single "background" class. Modeling the interactions between data and labels separately, as (1) does, is often too limiting for many computer vision tasks; we therefore use a recently proposed model that handles the interaction between site labels in a context-dependent way [9], describing it next.

The random field graph topology commonly used for joint image labeling problems is the lattice, (Figure 1), where cliques are single nodes and edges. We use a homogeneous, anisotropic random field. Thus, cliques of the same class use the same compatibility functions regardless of image location, but horizontal and vertical edges are considered different classes and thus have distinct compatibility functions. Anisotropy allows the model to learn any orientational bias of the labels. Our conditional random field has the form

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}) &= \frac{1}{Z(\mathbf{x})} \prod_{v \in V} \psi_V(y_v, \mathbf{x}) \prod_{(u,v) \in E} \psi_E(y_u, y_v, \mathbf{x}) & (2) \\ &= \frac{1}{Z(\mathbf{x})} \exp\left( \sum_{v \in V} \boldsymbol{\lambda} \cdot F(y_v, \mathbf{x}) + \sum_{(u,v) \in E} \boldsymbol{\mu} \cdot G(y_u, y_v, \mathbf{x}) \right), & (3) \end{aligned}$$

where $Z$ is now an observation-dependent normalizer. The compatibilities are functions of clique labels, allowing neighboring label interaction, but they are also functions of the entire observation. This differs markedly from (1) by allowing data-dependent label interaction (see Figure 1). $F$ and $G$ are vector-valued feature functions, and $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are vectors of parameters for nodes and edges, respectively. Node labels come from a discrete, finite alphabet $\mathcal{Y}$. We use one set of observation features for nodes and edges and transform them into feature functions (observation, label pairs) using the relationship

$$\begin{aligned} f_y^k(y_v, \mathbf{x}) &= \delta(y, y_v) f^k(\mathbf{x}) \\ g_{y,y'}^j(y_u, y_v, \mathbf{x}) &= \delta(y, y_u)\, \delta(y', y_v)\, g^j(\mathbf{x}), \end{aligned}$$

where $\mathbf{f} = \left(f^k\right)_{k=1\ldots K}$ is a vector of node features (i.e., texture statistics of a region) and $\mathbf{g} = \left(g^j\right)_{j=1\ldots J}$ is a vector of edge features (i.e., differences between statistics of neighboring regions), so that $F = \left(f_y^k\right)_{k=1\ldots K, y\in\mathcal{Y}}$ and $G = \left(g_{y,y'}^j\right)_{j=1\ldots J, y,y'\in\mathcal{Y}\times\mathcal{Y}}$. Thus $\boldsymbol{\lambda} \in \mathbb{R}^{K|\mathcal{Y}|}$ and $\boldsymbol{\mu} \in \mathbb{R}^{J|\mathcal{Y}|^2}$. When $E = \emptyset$, the model uses no label context and is commonly called a conditional maximum entropy classifier (hence, MaxEnt), or logistic regression.

**Training and Inference**

Parameters for probabilistic models like CRFs are generally set by maximizing the likelihood of a data sample. Unfortunately, inference for any random field with the lattice topology is intractable due to $Z$ (an exponential sum). Markov chain Monte Carlo (MCMC) (see e.g., [16]) is often used to approximate $Z$ in similar generative models. However, in our conditional model $Z$ is dependent on the image data $\mathbf{x}$ and must be estimated *for each observation* in the sample. A simpler approximation is to maximize the pseudo-likelihood (PL) [1], which is the product of the probabilities of nodes given their neighboring labels. The normalizers are then summations over labels at a single node, rather than the possible labelings of all nodes.

A relatively new alternative to MCMC and PL for approximating likelihood is called tree reparameterization (TRP) [15]. Inference in graphical models without cycles (unlike the lattice) is very efficient, i.e., due to the junction tree algorithm (e.g., [10]). An important consequence of the junction tree algorithm is that marginal distributions are revealed on pairs of neighboring nodes, inducing an alternative factorization of the joint distribution. TRP operates by using junction tree to compute the exact marginals on a spanning tree of the cyclic graph. The spanning tree's factorization is then placed back into the original graph, and the process repeats with different spanning trees until the parameterization converges, leaving the marginals. We demonstrate improved detection performance using TRP to approximate the likelihood over pseudo-likelihood. The likelihood function is convex and may be optimized globally via gradient ascent. Pseudo-likelihood is sensitive to initialization, however, so node parameters are optimized first. We use the quasi-newton L-BFGS algorithm for maximization.

To prevent training procedures from overfitting parameters in conditional models, a prior is introduced, and the posterior is maximized rather than likelihood. We employ a diagonal zero-centered Gaussian prior on parameters [2] (similar to weight decay in neural networks or ridge regression); variances are experimentally determined through cross-validation.

Given the image data, our model simply yields a joint posterior distribution on labelings. When interested in picking a hard and fast label for each region of the image (node in the graph), the question becomes what to do with that distribution. A simple, oft-used answer is to find its maximum. That is, use *maximum a posteriori* (MAP) estimation:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}^{|V|}} p\left(\mathbf{y} \mid \mathbf{x}\right).$$

This search space is intractable. However, a slight alteration of TRP allows MAP estimates to be quickly calculated. A simpler alternative is to search for a local maximum of the posterior, an estimate called iterated conditional modes (ICM). Given some initial labeling $\mathbf{y}^0$, subsequent labels are given by

$$y_v^{k+1} = \arg\max_{y_v \in \mathcal{Y}} p\left(y_v \mid y_{\mathcal{N}(v)}^k, \mathbf{x}\right), \forall v \in V$$

until $\mathbf{y}^{k+1} = \mathbf{y}^k$ or an iteration limit is exceeded. Often, the initial labeling comes from the local compatibility maximum $y_v^0 = \arg\max_{y_v \in \mathcal{Y}} \psi\left(y_v, \mathbf{x}\right)$. Like many point estimates, the MAP estimation has an important caveat: poor predictions can result when the maximum of the posterior is not representative of most of the other likely labelings [6]. An alternative method for prediction is called maximum posterior marginal (MPM) estimation,

$$\hat{y}_v = \arg\max_{y_v \in \mathcal{Y}} p\left(y_v \mid \mathbf{x}\right), \forall v \in V,$$

which accounts for the probability of all labelings, not simply the maximal (joint) labeling, by choosing the label at each node that maximizes its marginal probability. MAP and MPM are equivalent in the MaxEnt classifier since node labels are independent. Marginalization suffers from the same computational complexity problems as MAP, but since TRP reveals (approximate) marginals on the nodes, it is easily used for MPM. Comparisons between ICM and MAP estimated with TRP are given in the experiments.

## IMAGE FEATURES FOR SIGN DETECTION

Text and sign detection has been the subject of much research. Earlier approaches either use independent, local classifications (i.e., [5, 7, 11]) or use heuristic methods, such as connected component analysis (i.e., [4, 14]). Much work has been based on edge detectors or more general texture features, as well as color. Our approach calculates a joint labeling of image patches, rather than labeling patches independently, and it obviates layout heuristics by allowing the CRF to learn the characteristics of regions that contain text. Rather than simply using functions of single filters (e.g., moments) or edges, we use a richer representation that captures important relationships between responses to different scale- and orientation-selective filters.

To measure the general textural properties of both sign and especially non-sign (hence, background) image regions, we use responses of scale and orientation selective filters. Specifically, we use the statistics of filter responses described in [13], where correlations between steerable pyramid responses of different scales and orientations are the prominent features.

A biologically inspired non-linear texture operator for detecting gratings of bars at a particular orientation and scale is described in [12]. Scale and orientation selective filters, such as the steerable pyramid or Gabor filters, respond indiscriminately to both single edges and one or more bars. Grating

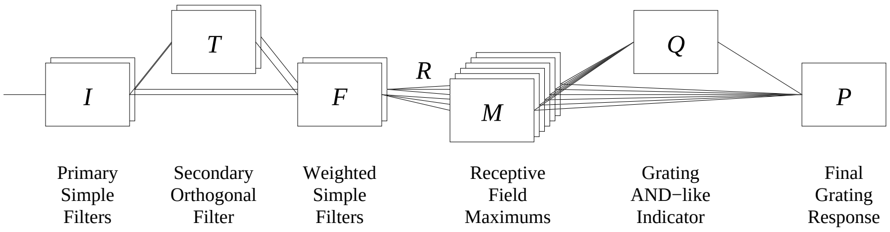


Figure 2: Grating cell data flow for a single scale and orientation. Two boxes at $I$, $T$, and $F$ represent center on and center off filters, while the boxes at $M$ are for the six receptive fields.

cells, on the other hand, respond selectively only to multiple (three or more) bars. This property is an ideal match for detecting text, which is generally characterized by a "grating" of strokes. The original model is contrast-normalized, but we expect text in signs to have high contrast for readability, so we omit any normalization when calculating $I_{\theta,\omega,\phi}$, the response of an input image to a filter with preferred orientation $\theta$, spatial frequency $\omega$, and phase $\phi$ (Figure 3, upper-right). Furthermore, letters have a limited aspect ratio, thus the bars in text have bounded height. For this reason we subject the responses $I_{\theta,\omega,\phi}$ to a second round of filtering with output $T_{\theta,\omega,\phi}$, where $\theta, \omega, \phi$ still indicates the parameters of the primary simple filter. The secondary filter has an orthogonal orientation $\theta + \frac{\pi}{2}$, a center-on phase of $\pi$, and should have a frequency of no more than $\omega/2$. To elicit stronger responses from bars of limited height, the original simple filter response is weighted by the perpendicular response with the Schur product $F_{\theta,\omega,\phi} \triangleq I_{\theta,\omega,\phi} \circ T_{\theta,\omega,\phi}$. Once the weighted responses are calculated, a binary grating cell subunit $Q_{\theta,\omega}$ indicates the presence of a grating at each image location. To make such a determination, alternating strong maximum center-on ($\phi = \pi$) and center-off ($\phi = 0$) responses $M_{\theta,\omega,n}$ are required in receptive field regions $\mathcal{R}_{\theta,\omega,n}$ ($-3 \leq n \leq 2$) of length $1/(2\omega)$ along a line with orientation $\theta$ (Figure 3, bottom). We let the final output $P_{\theta,\omega}$ be the mean response among the receptive fields where $Q_{\theta,\omega}$ indicates a grating and zero elsewhere. This also differs from the original model, which simply gives the spatial average of the grating indicator. Use of actual filter responses in the output is important because it represents the strength of the grating, rather than only its presence. After taking maximum responses over a set of scales, we use the mean, max, variance, skew and kurtosis of the outputs in a region as features.

Additionally, histograms of patch hue and saturation are used, which also allows us to measure color discontinuities between patches.

Using an algorithm [3] that ranks discriminative power of random field model features, we found the top three in the edge-less, context-free MaxEnt model to be (i) the level of green hue (easily identifying vegetation as background), (ii) mean grating cell response (easily identifying text), and (iii) correlation between a vertically and diagonally oriented filter of moderate scale (the single most useful other 'textural' feature).


Figure 3: Grating operator on text. UPPER LEFT: Input image. UPPER RIGHT: Center-on and center-off simple filter responses ($\theta = 0$). BOTTOM: Slice of simple filter responses and receptive regions for a marked point.


Figure 4: Multi-scale text detection with grating cells. LEFT: Input image with sign areas outlined. RIGHT: Grating cell responses.

## EXPERIMENTS

Our sign experiments are based on a hand-labeled database of 309 images collected from a North American downtown area with a still camera.[1] We view the 1024x768 pixel images as an 8x6 grid of 128x128 pixel patches over which the features are computed. This outer scale was chosen to balance computational burden against typical sign size; some patches contain more sign than others. Let $\mathbf{f}_p$ represent the statistics of the steerable pyramid, $\mathbf{f}_g$ the grating statistics, and $\mathbf{f}_h$, $\mathbf{f}_s$ the hue and saturation histograms, respectively. Our node features are the concatenated vectors $\mathbf{f} = \langle \mathbf{f}_p, \mathbf{f}_g, \mathbf{f}_h, \mathbf{f}_s, 1 \rangle$

[1] Available at `<http://vis-www.cs.umass.edu/projects/vidi>`.

| Classifier | | Prediction | Recall | Precision | F1 |
|---|---|---|---|---|---|
| MaxEnt | | MAP | 48.36 | 68.02 | 56.45 |
| CRF | PL | ICM | 49.42 | 70.23 | 57.90 |
| | | MAP | 49.53 | 70.23 | 57.97 |
| | | MPM | 49.97 | 69.77 | 58.13 |
| | TRP | ICM | 54.01 | 66.54 | 59.49 |
| | | MAP | 54.58 | 66.07 | 59.57 |
| | | MPM | 54.58 | 66.07 | 59.65 |

Table 1: Prediction results for signs. PL indicates training with pseudo-likelihood, and TRP training with approximated full likelihood. MAP and MPM for the CRF is estimated with TRP.

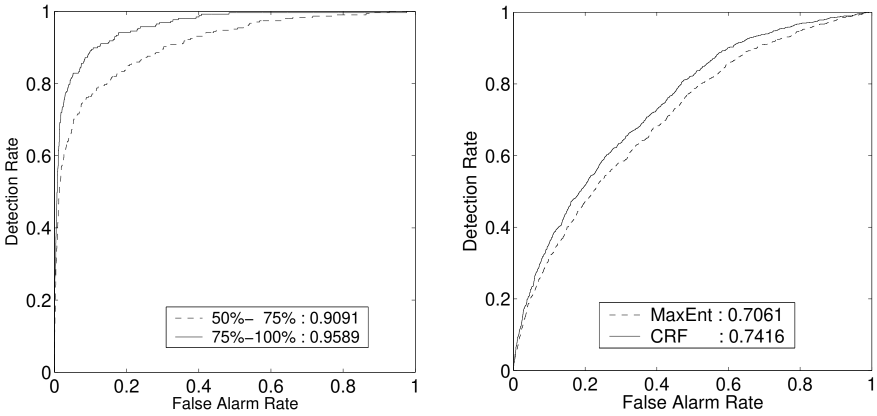


Figure 5: Discriminative power. Left: ROC curves and areas for MaxEnt on patches that are nearly all (75%-100%) sign or mostly (50-75%) sign. Right: ROC curves and areas for MaxEnt and CRF in patches containing less than 25% sign.

plus a bias feature. Edge features are the $L2$ norms of differences between statistics at neighboring patches, $\mathbf{g} = \left\langle \left\|\mathbf{f}_p - \mathbf{f}'_p\right\|, \|\mathbf{f}_h - \mathbf{f}'_h\|, \|\mathbf{f}_s - \mathbf{f}'_s\|, 1\right\rangle$.

The image set is split evenly with half each for training and testing. Table 1 contains the average prediction results of 20 such splits. Since this is a detection task, we report precision and recall (common in information retrieval) for each prediction method. Let $D_S$ be the number of true sign patches detected with $D$ the total number of detections and $S$ the actual number of true sign patches. Precision is $P = D_S/D$, the percentage of detections that are correct (not the complement of false alarm). Recall is $R = D_S/S$, the detection rate. The harmonic mean of recall and precision $F1 = 2PR/(P+R)$ reflects the balance (or lack thereof) between the rate and accuracy of detections; higher F1 indicates better overall performance.

MAP and ICM are point estimates of the unwieldy joint posterior probability, but the marginal posterior of a label (i.e., "sign") at a node is a real quantity that may be easily varied. Figure 5 (left) demonstrates that overall discrimination is very good even in the context-free MaxEnt classifier when

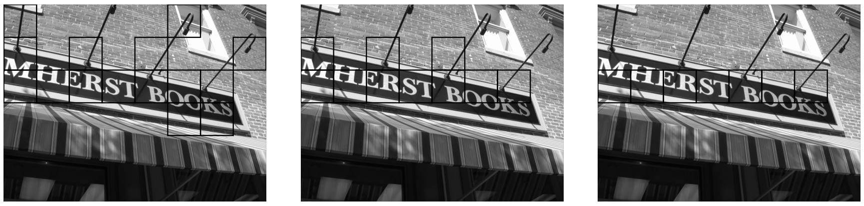

Figure 6: Example detection results. Left-Right: MaxEnt, CRF ICM and MAP.

a patch contains nearly all sign, but performance degrades as the amount of sign in a patch decreases. Figure 5 (right) shows that adding context with a CRF improves the ability to identify all regions of a sign, especially those on the border where the patch contains more background.

Using a CRF significantly improves F1 over the local MaxEnt classifier.[2] Training with TRP also improves recall. Because it is given true neighboring labels, which are unavailable at test time, PL training tends to be overconfident with edge parameters, leading to higher precision (excepting MPM) as a result of over-smoothing the labels. TRP training yields higher F1 and recall over PL for all prediction methods.

## CONCLUSIONS

The conditional random field is a powerful new model for vision applications that does not require the strong independence assumptions of generative models. With it, we demonstrate sign detection in natural images using both general texture features and special features for text. Adding context increases the detection rate faster than the false alarm rate by drawing on both observed data and unknown labels from neighboring regions.

The complexity issues of cyclic random fields are well known. Although training times are greater, prediction with a CRF still only requires about 3 seconds on a 3GHz desktop workstation. We have shown the superiority of tree reparameterization over the pseudo-likelihood approximation for parameter estimation and prediction in the CRF model for our detection task.

We plan to add more edge features to increase our use of the model's contextual power by incorporating feature selection and induction methods. Overfitting remains a constant problem in such a high-dimensional model, so regularization is an important area for study.

### Acknowledgments

Thanks to Aron Culotta, Khashayar Rohanimanesh, and Charles Sutton for their assistive discussions. This work was supported in part by NFS grant #IIS-0100851, in part by the Center for Intelligent Information Retrieval, and in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249.

[2]Claims of relative performance are based on a two-sided, paired sign test ($p < 4\mathrm{e}-5$).

## References


[1] J. Besag, "Statistical analysis of non-lattice data," **The Statistician**, vol. 24, no. 3, pp. 179–195, 1975.

[2] S. Chen and R. Rosenfeld, "A Gaussian prior for smoothing maximum entropy models," Techn. Report CMU-CS-99-108, **Carnegie Mellon University**, 1999.

[3] S. Della Pietra, V. Della Pietra and J. Lafferty, "Inducing Features of Random Fields," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 19, no. 4, pp. 380–393, 1997.

[4] J. Gao and J. Yang, "An Adaptive Algorithm for Text Detection from Natural Scenes," in **Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition**, December 2001, vol. 2, pp. 84–89.

[5] C. Garcia and X. Apostolidis, "Text Detection and Segmentation in Complex Color Images," in **Proceedings of 2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2000)**, June 2000, vol. 4, pp. 2326–2330.

[6] D. Grieg, B. Porteous and A. Seheult, "Exact maximum a posteriori estimation for binary images," **Journal of the Royal Statistical Society**, vol. 51, no. 2, pp. 271–279, 1989.

[7] A. Jain and S. Bhattacharjee, "Text segmentation using Gabor filters for automatic document processing," **Machine Vision Applications**, vol. 5, pp. 169–184, 1992.

[8] S. Kumar and M. Hebert, "Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification," in **Proc. 2003 IEEE International Conference on Computer Vision (ICCV '03)**, 2003, vol. 2, pp. 1150–1157.

[9] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in **Proc. 18th International Conference on Machine Learning**, Morgan Kaufmann, San Francisco, CA, 2001, pp. 282–289.

[10] S. L. Lauritzen, **Graphical Models**, no. 17 in Oxford Statistical Science Series, Clarendon, 1996.

[11] H. Li, D. Doermann and O. Kia, "Automatic Text Detection and Tracking in Digital Video," **IEEE Transactions on Image Processing**, vol. 9, no. 1, pp. 147–156, 2000.

[12] N. Petkov and P. Kruizinga, "Computational model of visual neurons specialised in the detection of period and aperiodic oriented visual stimuli: bar and grating cells," **Biological Cybernetics**, vol. 76, pp. 83–96, 1997.

[13] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," **International Journal of Computer Vision**, vol. 40, no. 1, pp. 49–71, 2000.

[14] R. M. Victor Wu and E. M. Riseman, "Finding Text in Images," in **DL'97: Proceedings of the 2nd ACM International Conference on Digital Libraries, Images, and Multimedia**, 1997, pp. 3–12.

[15] M. Wainwright, T. Jaakkola and A. Willsky, "Tree-based reparameterization framework for analysis of sum-product and related algorithms," **IEEE Transactions on Image Processing**, vol. 49, no. 5, pp. 1120–1146, 2003.

[16] G. Winkler, **Image Analysis, Random Fields, and Markov Chain Monte Carlo Methods**, Berlin: Springer-Verlag, 2nd edn., 2003.