Structured Learning and Prediction in Computer Vision

Structured Learning and Prediction in Computer Vision

Sebastian Nowozin

Microsoft Research Cambridge
United Kingdom
Sebastian.Nowozin@microsoft.com

Christoph H. Lampert

IST Austria
Institute of Science and Technology Austria
Austria
chl@ist.ac.at

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is S. Nowozin and C. H. Lampert, Structured Learning and Prediction in Computer Vision, Foundations and Trends® in Computer Graphics and Vision, vol 6, nos 3–4, pp 185–365, 2010

ISBN: 978-1-60198-456-2

Foundations and Trends® in Computer Graphics and Vision
Volume 6 Issues 3–4, 2010

Editorial Board

Editorial Scope

Foundations and Trends® in Computer Graphics and Vision will publish survey and tutorial articles in the following topics:

- Rendering: Lighting models; Forward rendering; Inverse rendering; Image-based rendering; Non-photorealistic rendering; Graphics hardware; Visibility computation
- Shape: Surface reconstruction; Range imaging; Geometric modelling; Parameterization;
- Mesh simplification
- Animation: Motion capture and processing; Physics-based modelling; Character animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape Representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and Video Retrieval
- Video analysis and event recognition
- Medical Image Analysis
- Robot Localization and Navigation

Information for Librarians
Foundations and Trends® in Computer Graphics and Vision, 2010, Volume 6, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Foundations and Trends® in
Computer Graphics and Vision
Vol. 6, Nos. 3–4 (2010) 185–365

DOI: 10.1561/0600000033

now
the essence of knowledge

Structured Learning and Prediction in Computer Vision

Sebastian Nowozin[1] and Christoph H. Lampert [2]

[1] *Microsoft Research Cambridge, United Kingdom, Sebastian.Nowozin@microsoft.com*
[2] *IST Austria, Institute of Science and Technology Austria, Austria, chl@ist.ac.at*

Abstract

Powerful statistical models that can be learned efficiently from large amounts of data are currently revolutionizing computer vision. These models possess a rich internal structure reflecting task-specific relations and constraints. This monograph introduces the reader to the most popular classes of structured models in computer vision. Our focus is discrete undirected graphical models which we cover in detail together with a description of algorithms for both probabilistic inference and maximum a posteriori inference. We discuss separately recently successful techniques for prediction in general structured models. In the second part of this monograph we describe methods for parameter learning where we distinguish the classic maximum likelihood based methods from the more recent prediction-based parameter learning methods. We highlight developments to enhance current models and discuss kernelized models and latent variable models. To make the monograph more practical and to provide links to further study we provide examples of successful application of many methods in the computer vision literature.

Contents

1

Introduction

In a very general sense *computer vision* is about automated systems making sense of image data by extracting some high-level information from it. The *image data* can come in a large variety of formats and modalities. It can be a single natural image, or it can be a multi-spectral satellite image series recorded over time. Likewise, the *high-level information* to be recovered is diverse, ranging from physical properties such as the surface normal at each image pixel to object-level attributes such as its general object class ("car," "pedestrian," etc.).

The above task is achieved by building a *model* relating the image data to the high-level information. The model is represented by a set of variables that can be divided into the *observation variables* describing the image data, the *output variables* defining the high-level information, and optionally a set of additional *auxiliary variables*. Besides the variables a model defines how the variables *interact* with each other. Together the variables and interactions form the *structure* of the model.

Structured models allow a large number of variables and interactions, leading to rich models that are able to represent the complex relationships that exist between the image data and the quantities of interest.

Instead of specifying a single-fixed model we can also introduce free *parameters* into the interactions. Given some annotated data with known values for the output variables we can then adjust the parameters to effectively *learn* a good mapping between observation and output variables. This is known as *parameter learning* and *training the model*.

1.1 An Example: Image Segmentation

We will now use the task of foreground–background image segmentation to make concrete the abstract concepts just discussed. In foreground–background image segmentation we are given a natural image and need to determine for each pixel whether it represents the foreground object or the background. To this end we define one binary output variable $y_i \in \{0,1\}$ for each pixel i, taking $y_i = 1$ if i belongs to the foreground, $y_i = 0$ otherwise. A single observation variable $x \in \mathcal{X}$ will represent the entire observed image.

To define the interactions between the variables we consider the following: if the image around a pixel i looks like a part of the foreground object, then $y_i = 1$ should be preferred over $y_i = 0$. More generally we may assume a local model $g_i(y_i, x)$, where $g_i(1, x)$ takes a high value if x looks like a foreground object around pixel i, and a low value otherwise. If this were the only component of the model we would make independent decisions for each pixel. But this is clearly insufficient. For example the model g_i might be inaccurate or the image locally really does resemble the foreground object. Therefore we introduce an interaction aimed at making locally consistent decisions about the output variables: for each pair (i, j) of pixels that are close to each other in the image plane — say within the 4-neighborhood $\mathcal{J}$ — we introduce a pairwise interaction term $g_{i,j}(y_i, y_j)$ that takes a large value if $y_i = y_j$ and a small value otherwise.

We can now pose segmentation as a maximization problem over all possible segmentations on n pixels,

$$y^* = \operatorname*{argmax}_{y \in \{0,1\}^n} \left[\sum_{i=1}^{n} g_i(y_i, x) + \sum_{(i,j) \in \mathcal{J}} g_{i,j}(y_i, y_j) \right]. \tag{1.1}$$

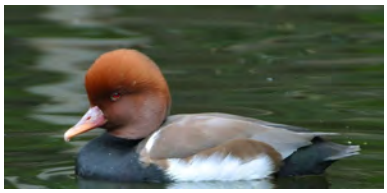

Fig. 1.1 Input image to be segmented into foreground and background. (Image source: `http://pdphoto.org`).

Fig. 1.2 Pixelwise separate classification by g_i only: noisy, locally inconsistent decisions.

Fig. 1.3 Joint optimum y^* with spatially consistent decisions.

The optimal prediction y^* will trade off the quality of the local model g_i with making decisions that are spatially consistent according to $g_{i,j}$. This is shown in Figures 1.1 to 1.3.

We did not say how the functions g_i and $g_{i,j}$ can be defined. In the above model we would use a simple binary classification model

$$g_i(y_i, x) = \langle w_{y_i}, \varphi_i(x) \rangle, \tag{1.2}$$

where $\varphi_i : \mathcal{X} \to \mathbb{R}^d$ extracts some image features from the image around pixel i, for example color or gradient histograms in a fixed window around i. The parameter vector $w_y \in \mathbb{R}^d$ weights these features. This allows the local model to represent interactions such as "if the picture around i is green, then it is more likely to be a background pixel." By adjusting $w = (w_0, w_1)$ suitably, a local score $g_i(y_i, x)$ can be computed for any given image. For the pairwise interaction $g_{i,j}(y_i, y_j)$ we ignore

the image x and use a 2×2 table of values for $g_{i,j}(0,0)$, $g_{i,j}(0,1)$, $g_{i,j}(1,0)$, and $g_{i,j}(1,1)$, for all adjacent pixels $(i,j) \in \mathcal{J}$.

1.2 Outline

In *Graphical Models* we introduce an important class of discrete structured models that can be concisely represented in terms of a graph. In this and later parts we will use *factor graphs*, a useful special class of graphical models. We do not address in detail the important class of *directed* graphical models and temporal models.

Computation in undirected discrete factor graphs in terms of probabilities is described in *Inference in Graphical Models*. Because for most models exact computations are intractable, we discuss a number of popular approximations such as belief propagation, mean field, and Monte Carlo approaches.

In *Structured Prediction* we generalize prediction with graphical models to the general case where a prediction is made by maximizing an arbitrary evaluation function, i.e., $y = f(x) = \text{argmax}_y g(x,y)$. Solving this problem — that is, evaluating $f(x)$ — is often intractable as well and we discuss general methods to approximately make predictions.

After having addressed these basic inference problems we consider learning of structured models. In *Conditional Random Fields* we introduce popular learning methods for graphical models. In particular we focus on recently proposed efficient methods able to scale to large training sets.

In *Structured Support Vector Machines* we show that learning is also possible in the general case where the model does not represent a probability distribution. We describe the most popular techniques and discuss in detail the structured support vector machine.

Throughout the monograph we interleave the main text with successful computer vision applications of the explained techniques. For convenience the reader can find a summary of the notation used at the end of the monograph.

References

[1] E. H. L. Aarts, J. H. M. Korst, and P. J. M. v. Laarhoven, "Simulated annealing," in *Local Search in Combinatorial Optimization*, (E. H. L. Aarts and J. K. Lenstra, eds.), pp. 91–120, Wiley-Interscience, 1997.

[2] R. K. Ahuja, Ö. Ergun, J. B. Orlin, and A. P. Punnen, "A survey of very large-scale neighborhood search techniques," in *Workshop on Discrete Optimization*, (E. Boros and P. L. Hammer, eds.), pp. 75–102, 2002.

[3] K. Alahari, P. Kohli, and P. H. S. Torr, "Reduce, reuse & recycle: Efficiently solving multi-label MRFs," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] K. Alahari, C. Russell, and P. H. S. Torr, "Efficient piecewise learning for conditional random fields," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[6] K. M. Anstreicher and L. A. Wolsey, "Two "well-known" properties of subgradient optimization," *Mathematical Programming*, vol. 120, no. B, pp. 213–220, 2009.

[7] A. U. Asuncion, Q. Liu, A. T. Ihler, and P. Smyth, "Learning with blocks: Composite likelihood and contrastive divergence," in *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2010.

[8] F. Barahona and R. Anbil, "The volume algorithm: Producing primal solutions with a subgradient method," *Mathematical Programming*, vol. 87, no. 3, pp. 385–399, 2000.

[9] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2011. In press.

[10] A. Barbu and S. C. Zhu, "Generalizing swendsen-wang to sampling arbitrary posterior probabilities," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 27, no. 8, pp. 1239–1253, 2005.

[11] D. Batra, A. C. Gallagher, D. Parikh, and T. Chen, "Beyond trees: MRF inference via outer-planar decomposition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[12] D. Batra, S. Nowozin, and P. Kohli, "Tighter relaxations for MAP-MRF inference: A local primal-dual gap based separation algorithm," in *Conference on Uncertainty in Artificial Intelligence (AISTATS)*, 2011.

[13] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 2nd Edition, 1995.

[14] D. P. Bertsekas, *Network Optimization*. Athena Scientific, 1998.

[15] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Athena Scientific, 1997.

[16] J. Besag, "Statistical analysis of non-lattice data," *The Statistician*, pp. 179–195, 1975.

[17] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society*, vol. B-48, no. 3, pp. 259–302, 1986.

[18] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 20, no. 4, pp. 401–406, 1998.

[19] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[20] A. Blake, C. Rother, M. Brown, P. Perez, and P. H. S. Torr, "Interactive image segmentation using an adaptive GMMRF model," in *European Conference on Computer Vision (ECCV)*, pp. 428–441, 2004.

[21] M. B. Blaschko and C. H. Lampert, "Learning to localize objects with structured output regression," in *European Conference on Computer Vision (ECCV)*, Springer, 2008.

[22] L. Bo and C. Sminchisescu, "Structured output-associative regression," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[23] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical Optimization*. Springer, 2003.

[24] A. Bordes, L. Bottou, and P. Gallinari, "SGD-QN: Careful Quasi-Newton stochastic gradient descent," *Journal of Machine Learning Research (JMLR)*, vol. 10, pp. 1737–1754, 2009.

[25] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Conference on Neural Information Processing Systems (NIPS)*, The MIT Press, 2007.

[26] Y. Boykov and V. Kolmogorov, "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision," *PAMI*, vol. 26, no. 9, pp. 1124–1137, 2004.

[27] Y. Boykov, O. Veksler, and R. Zabih, "Markov Random Fields with Efficient Approximations," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 648–655, IEEE Computer Society, 1998.

[28] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 23, no. 11, pp. 1222–1239, 2001.

[29] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *International Conference on Computer Vision (ICCV)*, pp. 105–112, 2001.

[30] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3241–3248, 2010.

[31] C. Chen, D. Freedman, and C. H. Lampert, "Enforcing topological constraints in random field image segmentation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2089–2096, 2011.

[32] J. Clausen, *Branch and Bound Algorithms — Principles and Examples*. University of Copenhagen, 1999.

[33] M. Collins, "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms," in *Conference on Empirical methods in Natural Language Processing*, pp. 1–8, 2002.

[34] A. J. Conejo, E. Castillo, R. Mínguez, and R. García-Bertrand, *Decomposition Techniques in Mathematical Programming*. Springer, 2006.

[35] K. Crammer and Y. Singer, "Ultraconservative online algorithms for multiclass problems," *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 951–991, 2003.

[36] G. Elidan, I. McGraw, and D. Koller, "Residual belief propagation: Informed scheduling for asynchronous message passing," in *Uncertainty in Artificial Intelligence (UAI)*, 2006.

[37] P. Felzenszwalb and D. Huttenlocher, "Efficient matching of pictorial structures," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 66–75, 2000.

[38] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision (IJCV)*, vol. 70, no. 1, pp. 41–54, 2006.

[39] T. Finley and T. Joachims, "Training structural SVMs when exact inference is intractable," in *International Conference on Machine Learing (ICML)*, pp. 304–311, 2008.

[40] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Computer*, vol. 22, no. 1, pp. 67–92, January 1973.

[41] R. Fletcher, *Practical Methods of Optimization*. John Wiley & Sons, 1987.

[42] V. Franc, S. Sonnenburg, and T. Werner, "Cutting-plane methods in machine learning," in *Optimization for Machine Learning*, (S. Sra, S. Nowozin, and S. J. Wright, eds.), MIT Press, 2011.

[43] A. Frangioni, "About lagrangian methods in integer optimization," *Annals of Operations Research*, vol. 139, no. 1, pp. 163–193, 2005.

[44] D. Freedman and P. Drineas, "Energy minimization via graph cuts: Settling what is possible," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 939–946, 2005.

[45] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International Journal of Computer Vision (IJCV)*, vol. 40, no. 1, pp. 25–47, 2000.

[46] B. J. Frey and D. J. C. MacKay, "A revolution: Belief propagation in graphs with cycles," in *Conference on Neural Information Processing Systems (NIPS)*, The MIT Press, 1997.

[47] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *International Conference on Computer Vision (ICCV)*, 2009.

[48] D. Geiger and A. L. Yuille, "A common framework for image segmentation," *International Journal of Computer Vision (IJCV)*, vol. 6, no. 3, pp. 227–243, 1991.

[49] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 6, no. 6, pp. 721–741, 1984.

[50] A. M. Geoffrion, "Lagrangian relaxation for integer programming," *Mathematical Programming Study*, vol. 2, pp. 82–114, 1974.

[51] C. J. Geyer, "Practical Markov chain Monte Carlo," *Statistical Science*, vol. 7, no. 4, pp. 473–483, 1992.

[52] J. Goodman, "Exponential priors for maximum entropy models," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 305–312, 2004.

[53] M. Guignard, "Lagrangean relaxation," *TOP*, vol. 11, no. 2, pp. 151–200, 2003.

[54] M. Guignard and S. Kim, "Lagrangean decomposition: A model yielding stronger Lagrangean bounds," *Mathematical Programming*, vol. 39, pp. 215–228, 1987.

[55] O. Häggström, *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, 2000.

[56] P. Hart, N. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[57] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, pp. 97–109, 1970.

[58] X. He, R. Zemel, and M. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[59] T. Heskes, "Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies," *Journal of Artificial Intelligence Research (JAIR)*, vol. 26, pp. 153–190, 2006.

[60] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, no. 6, pp. 409–436, 1952.

[61] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[62] 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.

[63] T. Joachims, T. Finley, and C.-N. Yu, "Cutting-plane training of structural SVMs," *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.

[64] J. K. Johnson, D. M. Malioutov, and A. S. Willsky, "Lagrangian relaxation for MAP estimation in graphical models," *Allerton Conference on Control, Communication and Computing*, 2007.

[65] V. Jojic, S. Gould, and D. Koller, "Accelerated dual decomposition for MAP inference," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21–24, 2010, Haifa, Israel*, (J. Fürnkranz and T. Joachims, eds.), pp. 503–510, Omnipress, 2010.

[66] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[67] J. Kelley Jr, "The cutting-plane method for solving convex programs," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 4, pp. 703–712, 1960.

[68] B. M. Kelm, N. Müller, B. H. Menze, and F. A. Hamprecht, "Bayesian estimation of smooth parameter maps for dynamic contrast-enhanced MR images with block-ICM," in *CVPR Workshop on Mathematical Methods in Biomedical Image Analysis, Computer Vision and Pattern Recognition*, pp. 96–103, 2006.

[69] R. Kikuchi, "A Theory of Cooperative Phenomena," *Physical Review*, vol. 81, no. 6, pp. 988–1003, 1951.

[70] T. Kim, S. Nowozin, P. Kohli, and C. D. Yoo, "Variable grouping for energy minimization," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[71] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.

[72] J. Kittler and J. Föglein, "Contextual classification of multispectral pixel data," *Image Vision Computing*, vol. 2, no. 1, pp. 13–29, 1984.

[73] P. Kohli, M. P. Kumar, and C. Rother, "MAP inference in discrete models," Tutorial at ICCV 2009, 2009.

[74] P. Kohli, M. P. Kumar, and P. H. S. Torr, "P3 & beyond: Solving energies with higher order cliques," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[75] P. Kohli, L. Ladický, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[76] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *International Journal of Computer Vision (IJCV)*, vol. 82, no. 3, pp. 302–324, 2009.

[77] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[78] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 28, no. 10, pp. 1568–1583, 2006.

[79] V. Kolmogorov and C. Rother, "Minimizing nonsubmodular functions with graph cuts-A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 29, no. 7, pp. 1274–1279, 2007.

[80] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 26, no. 2, pp. 147–159, 2004.

[81] N. Komodakis and N. Paragios, "Beyond loose LP-relaxations: Optimizing MRFs by repairing cycles," in *European Conference on Computer Vision (ECCV)*, 2008.

[82] N. Komodakis, N. Paragios, and G. Tziritas, "MRF optimization via dual decomposition: Message-passing revisited," in *International Conference on Computer Vision (ICCV)*, 2007.

[83] N. Komodakis, G. Tziritas, and N. Paragios, "Fast, approximately optimal solutions for single and dynamic MRFs," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 2007.

[84] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*. Springer, 4th Edition, 2008.

[85] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[86] A. Kulesza and F. Pereira, "Structured learning with approximate inference," in *Conference on Neural Information Processing Systems (NIPS)*, 2007.

[87] S. Kumar and M. Hebert, "Discriminative fields for modeling spatial dependencies in natural images," *Conference on Neural Information Processing Systems (NIPS)*, 2004.

[88] C. H. Lampert and M. B. Blaschko, "Structured prediction by joint kernel support estimation," *Machine Learning*, vol. 77, no. 2-3, pp. 249–269, 2009.

[89] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by Efficient Subwindow Search," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[90] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient subwindow search: A branch and bound framework for object localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 31, no. 12, pp. 2129–2142, 2009.

[91] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 67–81, 2004.

[92] C. Lemaréchal, "Lagrangian relaxation," in *Computational Combinatorial Optimization*, (M. Jünger and D. Naddef, eds.), pp. 112–156, Springer, 2001.

[93] V. S. Lempitsky, A. Blake, and C. Rother, "Image segmentation by branch-and-mincut," in *European Conference on Computer Vision (ECCV)*, 2008.

[94] V. S. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *International Conference on Computer Vision (ICCV)*, 2009.

[95] F. Liang, C. Liu, and R. J. Carroll, *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. John Wiley, 2010.

[96] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1, pp. 503–528, 1989.

[97] J. S. Liu, *Monte Carlo Strategies in Scientific Computing, Springer Series in Statistics*. New York: Springer, 2001.

[98] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[99] A. F. T. Martins, N. A. Smith, and E. P. Xing, "Polyhedral outer approximations with application to natural language parsing," in *International Conference on Machine Learing (ICML)*, 2009.

[100] A. F. T. Martins, N. A. Smith, E. P. Xing, P. M. Aguiar, and M. A. T. Figueiredo, "Augmenting dual decomposition for MAP inference," in *Proceedings of the 3rd International Workshop on Optimization for Machine Learning (OPT 2010), December 10, 2010, Whistler, Canada*, 2010.

[101] D. McAllester, "Generalization bounds and consistency for structured labeling," in *Predicting Structured Data*, (G. Bakır, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, eds.), The MIT Press, 2007.

[102] T. Meltzer, A. Globerson, and Y. Weiss, "Convergent message passing algorithms — a unifying view," in *Uncertainty in Artificial Intelligence (UAI)*, 2009.

[103] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *et al.*, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[104] M. Mézard and A. Montanari, *Information, Physics and Computation*. Oxford University Press, 2009.

[105] T. Minka, "Divergence measures and message passing," Microsoft Research Technical Report, MSR-TR-2005-173, 2005.

[106] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[107] G. Mori, "Guiding model search using segmentation," in *International Conference on Computer Vision (ICCV)*, 2005.

[108] I. Murray, "Markov chain Monte Carlo," Tutorial at Machine Learning Summer School 2009, 2009.

[109] A. Nedic and D. Bertsekas, "Convergence rate of incremental subgradient algorithms," in *Stochastic Optimization: Algorithms and Applications*, (S. P. Uryasev and P. M. Pardalos, eds.), pp. 223–264, Springer, 2000.

[110] Y. E. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[111] J. Neter, M. Kutner, C. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*. McGraw-Hill, 4 Edition, 1996.

[112] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.
[113] S. Nowozin, P. V. Gehler, and C. H. Lampert, "On parameter learning in CRF-based approaches to object class image segmentation," in *European Conference on Computer Vision (ECCV)*, 2010.
[114] S. Nowozin and S. Jegelka, "Solution stability in linear programming relaxations: graph partitioning and unsupervised learning," in *International Conference on Machine Learing (ICML)*, 2009.
[115] S. Nowozin and C. H. Lampert, "Global connectivity potentials for random field models," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
[116] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: Algorithms and complexity*. Dover Publications, 1998.
[117] J. C. Picard and M. Queyranne, "On the structure of all minimum cuts in a network and applications," *Combinatorial Optimization II*, pp. 8–16, 1980.
[118] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*, (B. Schölkopf, C. J. C. Burges, and A. J. Smola, eds.), pp. 185–208, The MIT Press, 1999.
[119] J. C. Platt, "Probabilities for SV Machines," in *Advances in Large Margin Classifiers*, (A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, eds.), pp. 61–74, The MIT Press, 1999.
[120] P. Pletscher, C. S. Ong, and J. M. Buhmann, "Entropy and margin maximization for structured output learning," in *European Conference on Marchine Learning (ECML)*, 2010.
[121] B. Potetz, "Efficient belief propagation for vision using linear constraint nodes," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
[122] S. Ramalingam, P. Kohli, K. Alahari, and P. H. S. Torr, "Exact inference in multi-label CRFs with higher order cliques," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
[123] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *International Conference on Machine Learing (ICML)*, 2006.
[124] X. Ren and J. Malik, "Learning a classification model for segmentation," in *International Conference on Computer Vision (ICCV)*, 2003.
[125] C. P. Robert, *The Bayesian Choice. From Decision Theoretic Foundations to Computational Implementation, Springer Series in Statistics*. Springer, 2001.
[126] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2nd Edition, 2004.
[127] C. Rother, V. Kolmogorov, V. S. Lempitsky, and M. Szummer, "Optimizing binary MRFs via extended roof duality," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
[128] S. Sarawagi and R. Gupta, "Accurate max-margin training for structured output spaces," in *International Conference on Machine Learing (ICML)*, 2008.
[129] L. K. Saul and M. I. Jordan, "Exploiting tractable substructures in intractable networks," in *Conference on Neural Information Processing Systems (NIPS)*, pp. 486–492, 1995.

[130] B. D. Savchynskyy, J. Kappes, S. Schmidt, and C. Schnörr, "A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 2011.

[131] M. I. Schlesinger, "Syntactic analysis of two-dimensional visual signals in noisy conditions in Russian," *Kibernetika*, vol. 4, pp. 113–130, 1976.

[132] F. R. Schmidt, D. Farin, and D. Cremers, "Fast matching of planar shapes in sub-cubic runtime," in *International Conference on Computer Vision (ICCV)*, 2007.

[133] F. R. Schmidt, E. Töppe, and D. Cremers, "Efficient planar graph cuts with applications in computer vision," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[134] U. Schmidt, Q. Gao, and S. Roth, "A generative perspective on MRFs in low-level vision," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[135] B. Schölkopf and A. J. Smola, *Learning with Kernels*. The MIT Press, 2002.

[136] N. N. Schraudolph and D. Kamenetsky, "Efficient exact inference in planar ising models," in *Conference on Neural Information Processing Systems (NIPS)*, The MIT Press, 2008.

[137] N. N. Schraudolph and D. Kamenetsky, "Efficient exact inference in planar ising models," in *Conference on Neural Information Processing Systems (NIPS)*, 2008.

[138] S. E. Shimony, "Finding MAPs for belief networks Is NP-hard," *Artificial Intelligence*, vol. 68, no. 2, pp. 399–410, 1994.

[139] N. Z. Shor, *Minimization Methods for Non-differentiable Functions*. Springer, 1985.

[140] D. Sontag, A. Globerson, and T. Jaakkola, "Clusters and coarse partitions in LP relaxations," in *Conference on Neural Information Processing Systems (NIPS)*, 2008.

[141] D. Sontag, A. Globerson, and T. Jaakkola, "Introduction to dual decomposition for inference," in *Optimization for Machine Learning*, (S. Sra, S. Nowozin, and S. J. Wright, eds.), MIT Press, 2011.

[142] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss, "Tightening LP Relaxations for MAP using Message Passing," in *Uncertainty in Artificial Intelligence (UAI)*, pp. 503–510, 2008.

[143] P. Strandmark and F. Kahl, "Parallel and distributed graph cuts by dual decomposition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[144] C. Sutton and A. McCallum, "Piecewise training for structured prediction," *Machine Learning*, vol. 77, no. 2–3, pp. 165–194, 2009.

[145] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. F. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 30, no. 6, pp. 1068–1080, 2008.

[146] M. Szummer, P. Kohli, and D. Hoiem, "Learning CRFs using graph cuts," in *European Conference on Computer Vision (ECCV)*, 2008.

[147] M. F. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," in *International Conference on Computer Vision (ICCV)*, pp. 900–907, IEEE Computer Society, 2003.

[148] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Conference on neural information processing systems (NIPS)*, 2003.

[149] C. Teo, S. Vishwanathan, A. Smola, and Q. Le, "Bundle methods for regularized risk minimization," *Journal of Machine Learning Research (JMLR)*, vol. 1, pp. 1–55, 2009.

[150] L. Torresani, V. Kolmogorov, and C. Rother, "Feature correspondence via graph matching: Models and global optimization," in *European Conference on Computer Vision (ECCV)*, 2008.

[151] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research (JMLR)*, vol. 6, pp. 1453–1484, 2005.

[152] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition," *International Journal of Computer Vision (IJCV)*, vol. 63, no. 2, pp. 113–140, 2005.

[153] D. Tuia, J. Muñoz-Marí, M. Kanevski, and G. Camps-Valls, "Structured output SVM for remote sensing image classification," *Journal of Signal Processing Systems*, 2010.

[154] V. Vapnik and A. Chervonenkis, *Theory of Pattern Recognition (in Russian)*. Nauka, Moscow, 1974.

[155] V. V. Vazirani, *Approximation Algorithms*. Springer, 2001.

[156] A. Vedaldi and A. Zisserman, "Structured output regression for detection with partial occulsion," in *Conference on Neural Information Processing Systems (NIPS)*, 2009.

[157] S. Vicente, V. Kolmogorov, and C. Rother, "Graph cut based image segmentation with connectivity priors," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[158] S. Vicente, V. Kolmogorov, and C. Rother, "Joint optimization of segmentation and appearance models," in *International Conference on Computer Vision (ICCV)*, 2009.

[159] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *International Conference on Machine Learing (ICML)*, pp. 969–976, 2006.

[160] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[161] J. J. Weinman, L. Tran, and C. J. Pal, "Efficiently learning random fields for stereo vision with sparse message passing," in *European Conference on Computer Vision (ECCV)*, 2008.

[162] T. Werner, "A linear programming approach to max-sum problem: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 29, no. 7, pp. 1165–1179, 2007.

[163] T. Werner, "High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (MAP-MRF)," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[164] T. Werner, "Revisiting the decomposition approach to inference in exponential families and graphical models," Center for Machine Perception, Czech Technical University Prague, Research Report, CTU-CMP-2009-06, ftp://cmp.felk.cvut.cz/pub/cmp/articles/werner/Werner-TR-2009-06.pdf, 2009.

[165] T. Werner, "Belief propagation fixed points as zero gradients of a function of reparameterizations," Center for Machine Perception, Czech Technical University Prague, Research Report, CTU-CMP-2010-05, ftp://cmp.felk.cvut.cz/pub/cmp/articles/werner/Werner-TR-2010-05.pdf, 2010.

[166] H. P. Williams, *Model Building in Mathematical Programming*. New York: John Wiley & Sons, 4 Edition, 1999.

[167] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research (JMLR)*, vol. 6, pp. 661–694, 2005.

[168] L. A. Wolsey, *Integer Programming*. New York: John Wiley & Sons, 1998.

[169] O. J. Woodford, C. Rother, and V. Kolmogorov, "A global perspective on MAP inference for low-level vision," in *International Conference on Computer Vision (ICCV)*, 2009.

[170] E. P. Xing, M. I. Jordan, and S. J. Russell, "A generalized mean field algorithm for variational inference in exponential families," in *Uncertainty in Artificial Intelligence (UAI)*, pp. 583–591, 2003.

[171] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," MERL Technical Report, 2004-040, http://www.merl.com/papers/docs/TR2004-040.pdf, 2004.

[172] C. N. J. Yu and T. Joachims, "Learning structural SVMs with latent variables," in *International Conference on Machine Learing (ICML)*, 2009.

[173] A. Yuille, "The convergence of contrastive divergences," in *Conference on Neural Information Processing Systems (NIPS)*, pp. 1593–1600, 2005.

[174] A. L. Yuille, "CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation," *Neural Computation*, vol. 14, no. 7, pp. 1691–1722, 2002.

[175] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[176] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Annals of Statistics*, vol. 32, no. 1, pp. 56–85, 2004.

[177] S. C. Zhu, Y. N. Wu, and D. Mumford, "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling," *International Journal of Computer Vision (IJCV)*, vol. 27, no. 2, pp. 107–126, 1998.