MAP-Inference for Highly-Connected Graphs with DC-Programming

Jörg Kappes and Christoph Schnörr

Image and Pattern Analysis Group, Heidelberg Collaboratory for Image Processing,
University of Heidelberg, Germany,
{kappes,schnoerr}@math.uni-heidelberg.de

Abstract. The design of inference algorithms for discrete-valued Markov Random Fields constitutes an ongoing research topic in computer vision. Large state-spaces, none-submodular energy-functions, and highly-connected structures of the underlying graph render this problem particularly difficult. Established techniques that work well for sparsely connected grid-graphs used for image labeling, degrade for non-sparse models used for object recognition.
In this context, we present a new class of mathematically sound algorithms that can be flexibly applied to this problem class with a guarantee to converge to a critical point of the objective function. The resulting iterative algorithms can be interpreted as simple message passing algorithms that converge by construction, in contrast to other message passing algorithms.
Numerical experiments demonstrate its performance in comparison with established techniques.

1 Introduction

Applications of Markov Random Fields (MRFs) abound in computer vision. For a review and performance evaluation, we refer to [11].

The majority of applications amounts to some form of image labeling over sparsely connected grid graphs, akin to PDE-based processing in the continuous case. Established inference algorithms [12, 7] rely on convex relaxations and dedicated algorithms for solving the resulting large-scale linear programs (LPs). For a review, we refer to [13].

Yet, it has been recognized that the performance of these algorithms degrade for more involved problems that have a large number of states and highly-connected underlying graphs [6]. Such problems typically arise in connection with object recognition.

Another problem concerns convergence of the most attractive techniques. While it is well-known that loopy belief propagation is a heuristic from the viewpoint of algorithm design [14], sound relaxation techniques like tree-reweighted belief propagation may also suffer from convergence problems [12]. Techniques to remedy this [5] are based on restrictions that may not be satisfied in applications.

Finally, the quickly increasing number of constraints of LP-based relaxations is an issue caused by highly-connected graphs, in particular if the number of states is large, too. In this connection, Ravikumar and Lafferty [10] suggested recently a quadratic programming (QP) relaxation that essentially boils down to mean-field based MAP-estimation. Unlike the usual fixed-point iterations used in mean-field annealing, QP techniques can be applied, but the inherent non-convexity of this type of relaxation remains. To cope with it, a spectral rectification of the problem matrix was suggested in [10] in order to approximate the non-convex relaxation again by a convex one.

In this paper, we present a novel class of inference algorithms for MRF-inference, based on the non-convex relaxation introduced in [10]. This class is based on Difference of Convex Functions (DC) - programming that utilizes problem decompositions into two convex optimization problems. While in our field these technique has been previously applied for the marginalization problem [15], without referring to the vast and established mathematical literature [4], our present work applies these techniques for the first time to the MAP-inference problem, leading to fairly different algorithms.

We fix the notation and introduce the MAP-inference problem in section 2. The basic problem relaxation and the novel class of inference algorithms are detailed in Section 3, followed by a comparative numerical performance evaluation in Section 4.

2 Problem

2.1 Notation

Let $G = (V, E)$ be a graph with a set of nodes V and edges E. The variable x_s with $s \in V$ belongs to the set X_s, so the configuration space of all labellings x is $\bigotimes_{s \in V} X_s$. The costs assigned to any value of x are given by the model functions $\theta_{st}(x_s, x_t)$ and $\theta_s(x_s)\,,\ \forall s, t \in V\,,\ s \neq t$. The corresponding index sets are $\mathcal{I} = \mathcal{I}^V \cup \mathcal{I}^E$, $\mathcal{I}^V = \{(s; i) | s \in V, i \in X_s\}$ and $\mathcal{I}^E = \{(st; ij) | s, t \in V, i \in X_s, j \in X_t\}$.

2.2 MAP Inference

The maximum a posteriori (MAP) inference problem amounts to find a labeling x minimizing an energy function of the form

$$J(x) = \sum_{st \in E} \theta_{st}(x_s, x_t) + \sum_{s \in V} \theta_s(x_s)\ . \tag{1}$$

Assembling all function values of all terms into a single large vector $\theta \in \mathbb{R}^{|\mathcal{I}|}$, this problem can be shown to be equivalent to evaluating the support function of the marginal polytope[1] $\mathcal{M}$,

$$\sup_{\mu \in \mathcal{M}} \langle -\theta, \mu \rangle\ , \tag{2}$$

[1] The negative sign in (2) is due to our preference to work with energies that are to be *minimized*.

in terms of the vector of marginals μ. This problem, of course, is as intractable as is problem (1). But relaxations can be easily derived by replacing $\mathcal{M}$ by simpler sets. In this paper, we consider the simplest possibility, i.e. the product of all standard (probability) simplices over V:

$$\Lambda = \left\{ \mu \in \mathbb{R}_+^{\mathcal{I}^V} \,\middle|\, \sum_{i \in X_s} \mu_{s;i} = 1 \,,\ \forall s \in V \right\} \,. \tag{3}$$

3 Approach

3.1 QP-Relaxation

In [10], it was suggested to reduce the problem size by replacing $\mu \in \mathbb{R}_+^{|\mathcal{I}|}$ in (2) by a new set of variables $\tau \in \mathbb{R}_+^{|\mathcal{I}^V|}$. Inserting $\mu_{st;ij} = \tau_{s;i}\,\tau_{t;j}$ and $\mu_{s;i} = \tau_{s;i}$, problem (2) becomes a QP of a much smaller size

$$\begin{aligned} \min\ & \frac{1}{2}\tau^\top Q\tau + q^\top \tau \,, \\ s.t.\ & \tau \in \Lambda \,. \end{aligned} \tag{4}$$

This QP is not convex in general. Ravikumar and Lafferty [10] propose to base inference one a convex approximation, by adding a diagonal matrix in order to shift the spectrum of Q to the nonnegative cone, and by modifying the linear term accordingly, in view of extreme points of the set Λ,

$$\begin{aligned} \min\ & \frac{1}{2}\tau^\top (Q + diag(d))\tau + (q - \frac{1}{2}d)^\top \tau \,, \\ s.t.\ & \tau \in \Lambda \,. \end{aligned} \tag{5}$$

Good performance results with respect to inference are reported in [10], in addition to being computationally attractive for graphical models with large edge sets E due to the removal of many constraints. On the other hand, the original idea of inner-polytope relaxations in terms of a mean-field approximation to inference is inevitably lost through the convex approximation.

It is this latter problem that we address with a novel class of algorithms, to be described next.

3.2 DC-Decomposition

According to the discussion at the end of the previous section, we propose to dispense with convex approximations of (4), but to tackle it directly through DC-programming.

The basic idea is to decompose the non-convex symmetric quadratic part into the difference of two semi-definite quadratic forms

$$f(\tau) = \frac{1}{2}\tau^\top Q\tau + q^\top\tau = g(\tau) - h(\tau)\ , \tag{6a}$$

$$g(\tau) = \frac{1}{2}\tau^\top Q_1\tau + q^\top\tau\ , \tag{6b}$$

$$h(\tau) = -\frac{1}{2}\tau^\top Q_2\tau\ . \tag{6c}$$

Various choices of Q_1, Q_2 are possible:

Eigenvalue-Decomposition: DC_{EV}
Based on the spectral decomposition $Q = V diag(eig(Q))V^\top$, where V is the matrix of the eigenvectors and $eig(Q)$ are the eigenvalues of Q, we define

$$Q_1 = V diag(\max\{0, eig(Q)\}V^\top\ , \tag{7}$$

$$Q_2 = V diag(\max\{0, -eig(Q)\}V^\top\ . \tag{8}$$

Decomposition based on the Smallest Eigenvalue: DC_{MIN}
Since the computation of the eigenvalue decomposition is costly, another decomposition can be based on computing a lower bound for the smallest eigenvalue $d_{min} < \min\{eig(Q)\} < 0$. The smallest eigenvalue can be computed, e.g., by the power method.

$$Q_1 = Q - d_{min} \cdot I\ , \tag{9}$$

$$Q_2 = -d_{min} \cdot I\ . \tag{10}$$

Decomposition based on the Largest Eigenvalue: DC_{MAX}
A third immediate possibility utilizes the value of the largest eigenvalue of Q and has additionally the property that the convex part of the function becomes very simple. Let $d_{max} > \max\{eig(Q)\}$ be an upper bound of the largest eigenvalue. We define

$$Q_1 = d_{max} \cdot I\ , \tag{11}$$

$$Q_2 = d_{max} \cdot I - Q\ . \tag{12}$$

3.3 Inference Algorithm

For any decomposition introduced in the previous section, minimization is carried out by the general two-step iteration[2]

$$y^k \in \partial h(x^k)\ , \quad x^{k+1} \in \partial g^*(y^k)$$

that applies to any DC-objective $g(x) - h(x)$ and has known convergence properties [2]. Taking into account the specific structure of our objective function (6a), we arrive at the following simple algorithm: Depending on the particular decomposition $Q = Q_1 - Q_2$, the convex optimization steps 3 and 7 can be efficiently conducted with dedicated algorithms.

[2] $\partial f(x)$ denotes the set of subgradients of a proper convex lower-semicontinuous function f at x, and $g^*(y)$ denotes the Fenchel conjugate function $\sup_x\{\langle x, y\rangle - g(x)\}$

Algorithm 1 DC-Algorithm for MRF

[x] $\leftarrow$ **dc4mrf** (Q, q)

1: $[Q_1, Q_2] \Leftarrow \text{decompose}(Q)$
2: $i \Leftarrow 0$
3: $x^0 \Leftarrow \arg\min_x \frac{1}{2} x^\top Q_1 x + q^\top x \quad s.t.\ x \in \Lambda$ {Solve convex part}
4: **repeat**
5: $\quad i \Leftarrow i + 1$
6: $\quad y^{i-1} \Leftarrow Q_2 x^{(i-1)}$
7: $\quad x^i \Leftarrow \arg\min_x \frac{1}{2} x^\top Q_1 x + (q + y^{i-1})^\top x \quad s.t.\ x \in \Lambda$
8: **until** $\|x^i - x^{i-1}\|_\infty < \epsilon$

4 Experiments

We will compare nine different approaches to minimize the objective function given in (1) for full connected graphs. For computing the global optimum we use an A^*-based algorithm suggested in [3]. The three DC-decompositions, presented in this paper, are compared with the LP-relaxation [12], QP-relaxations [10], Belief Propagation (BP) [14], the TRBP-message passing algorithm by Wainwright [12] and the Second-Order-Cone-Programming (SOCP) approach by Kumar [9].

Since the QP-relaxation suggested in [10] turned out to be not tight for our data sets, we use $\lambda_{min} I$ to make the problem convex. For SOCP we do not use triangular constraints to keep the number of constraints manageable. Note, that the latest modifications [8] are not taken into account here, in order not to increase further the number of constraints. TRBP and BP are stopped if the largest change of a message is smaller than 10^{-6} or after 1000 iterations. A^*, BP and TRBP are C-implementations, DC_{MAX} runs with pure MatLab research code.

The constrained QPs achieved by DC_{MIN} and DC_{EV} as well as LP and QP, are solved using MOSEK [1]. Note, the constraints for DC_{MIN} and DC_{EV} are simple and performance can be increased using specific optimization-methods. We choose $\epsilon = 10^{-6}$ and set the maximal number of iterations for DC_{MAX} to 100000 and for DC_{MIN} and DC_{EV} to 1000. In cases where relaxations resulted in fractal solution we project it to the nearest integer solution.

Some approaches strongly depend to the number of constraints and variables, which limits the range of application. For instance, when using SOCP and LP, the number of constraints grow very fast. Table 1 summarizes the number of constraints and variables.

4.1 Synthetic Experiments

For synthetic experiments we generate graphical models and vary the number of nodes and the size of the state-space of each random variable. The potentials θ are sampled in three ways, given in Table 2.

Table 3 shows the mean energy, the calculation time (round brackets) and for experiment C the accuracy (square brackets). For experiments A and B our DC-

Table 1. Number of constraints and variables required for the different methods. L is the number of labels and K is the number of not truncated edges (see [9] for details). In the worst case $K \sim |E| \cdot L^2$

Method	Number of constraints	Number of variables
A^*	0	$\|V\|$
SOCP	$\|V\| + \|V\| \cdot L + 3K$	$\|V\| \cdot L + 2 \cdot K$
DC / QP	$\|V\|$	$\|V\| \cdot L$
LP	$\|V\| \cdot L + \|E\| \cdot L^2 + \|V\| \cdot L + 2 \cdot \|E\| \cdot L$	$\|V\| \cdot L + \|E\| \cdot L^2$
TRBP / BP	0	$2 \cdot \|E\| \cdot L$

Table 2. Overview of the sythetic experiments. In experiment 1 we compute uniform samples in 0 and 1 and set the potentials to the negative logarithm. In the second experiment we set all unary potentials to 0. In the third one we select a configuration x and set $\theta_{st} x_s, x_t$ and 5% or 10% entries in θ_{st} to 0 and the others to $-\log(0.1)$.

	Exp. A	Exp. B	Exp. C
θ_s	$-\log(\mathrm{U}(0,1))$	$-\log(1)$	$-\log(1)$
θ_{st}	$-\log(\mathrm{U}(0,1))$	$-\log(\mathrm{U}(0,1))$	$-\log(\{0.1,1\})$

approach outperforms BP, TRBP, SOCP, QP and LP. A^*, which always finds the global optimum, is only applicable to small graphs, of course. In experiment C, BP and TRBP are superior to our approach. However, there is no guarantee for convergence.

We also kept the ratio between nodes and labels fixed to 4 and did inference with the different approaches. The result is shown in Fig. 1. The plot for uniform sampled potentials (top line) again shows that DC_{MAX} outperforms state of the art inference techniques, at higher computational costs, however. Observe how for experiment C (bottom line) the run-time for SOCP and LP increases fast with the number of constraints.

To analyze the influence of the initial state to our DC-approach we selected x^0 in experiment A randomly, using 5 nodes and 20 labels. In the same experiments of Table 3, with 10 random initial states, we achieved a mean energy of 3.5878, which is fairly robust.

4.2 Application to Object Recognition

We tested different approaches on a real world scenario. The potential functions are now no longer generated synthetically, but estimated from real data through local classification.

We have two scenarios, the human face and the human body, as explained detailed in [3]. Due to the strong geometry, faces are simple and A^* solve this problems global optimal and fast. For human bodies A^* suffers from several shortcomings. For some images inference with A^* takes several minutes. The median with 0.21 seconds its still fast. In terms of energy we get the ranking $A^* < BP < DC_{MAX} < TRBP$ as visualized in Table 4, where the difference between BP an DC_{MAX} is not significant.

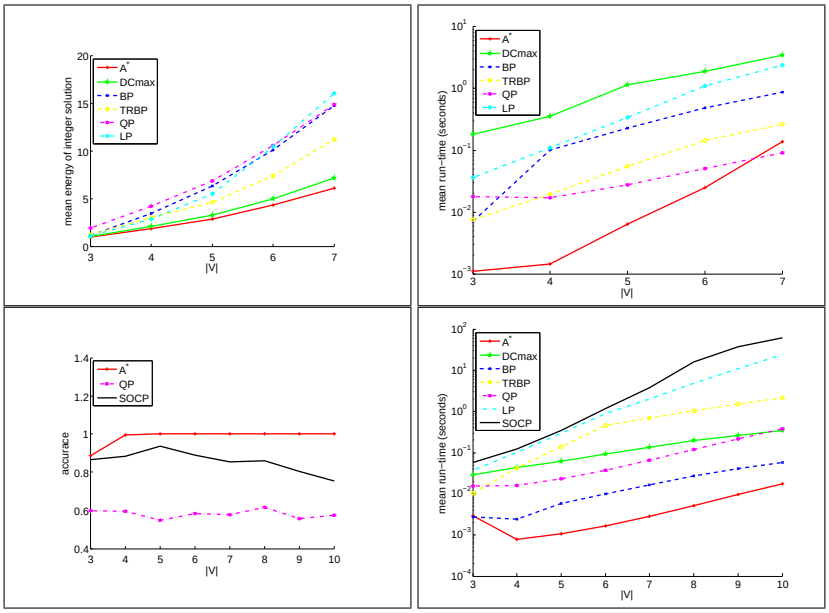

Fig. 1. In the first row we show energy- and runtime-plots for models of type A with 3 to 7 nodes. In the second row we show accuracy with respect to ground truth and runtime-plots for models of type C with 5% noise and 3 to 10 nodes. We use 4 times more labels than nodes. Accuracy for the other approaches are same than for A^*. DC_{MAX} finds in exp. A the lowest energy, next to A^*, and is comparable with state of the art approaches in exp C.

Fig. 2. visualize the three images with the largest relative energy of DC_{MAX}. From left to right the configurations of A^*, DC_{MAX}, BP and TRBP are shown. Surprisingly the DC_{MAX} solution matches the reality best.

Table 3. The table shows for different combinations of experiment settings and optimization techniques, the energy of integer solution, the required run-time (seconds) in round brackets and the accuracy in squared brackets. In the first column the numbers of variables and labels are given. We repeated each setting 100 times (5 times if marked with *) with random data. A dash indicates that the experiments could not be carried out, because the number of constraints is to high (LP, SOCP). For exp. A and B the DC-Algorithms outperforms state of the art approaches.

Experiment	A^*	DC_{MAX}	DC_{MIN}	DC_{EV}	BP	TRBP	SOCP	QP	LP
05-20-A	2.90	**3.29**	3.34	3.36	7.12	4.76	10.01	7.05	6.16
	(0.01)	(0.88)	(8.30)	(2.32)	(0.22)	(0.41)	(65.14)	(0.02)	(0.34)
05-20-B	1.07	**1.59**	1.61	1.68	6.30	4.86	6.39	5.43	6.04
	(0.01)	(1.89)	(10.20)	(2.29)	(0.28)	(0.11)	(65.21)	(0.02)	(0.38)
05-20-C-05%	0.00	**0.00**	**0.00**	0.62	**0.00**	**0.00**	1.82	13.22	**0.00**
	(0.01)	(0.06)	(0.56)	(0.11)	(0.01)	(0.14)	(0.33)	(0.02)	(0.29)
	[1.00]	[1.00]	[1.00]	[0.94]	[1.00]	[1.00]	[0.95]	[0.56]	[1.00]
05-20-C-10%	0.00	1.40	1.20	2.83	**0.00**	**0.00**	11.86	14.09	**0.00**
	(0.01)	(0.09)	(0.87)	(0.31)	(0.01)	(0.15)	(0.98)	(0.02)	(0.31)
	[1.00]	[0.80]	[0.83]	[0.64]	[1.00]	[1.00]	[0.46]	[0.38]	[1.00]
10-50-A	14.53*	**15.77**	15.90*	16.21*	42.35	30.03	-	33.75	36.76
	(1162)*	(21.59)	(1087)*	(808)*	(5.68)	(1.71)	-	(1.12)	(54.86)
10-50-B	9.75*	**12.11**	12.26*	12.36*	42.44	39.36	-	31.89	41.86
	(1297)*	(16.92)	(1109)*	(805)*	(5.74)	(1.68)	-	(1.08)	(53.49)
20-50-C-05%	0.00	**0.00**	**0.00***	**0.00***	**0.00**	**0.00**	-	47.94	-
	(0.02)	(1.21)	(4809)*	(5089)*	(0.33)	(9.83)	-	(5.08)	-
	[1.00]	[1.00]	[1.00]*	[1.00]*	[1.00]	[1.00]	-	[0.94]	-
20-50-C-10%	0.00	**0.00**	**0.00***	**0.00***	**0.00**	**0.00**	-	199.04	-
	(0.02)	(1.26)	(4707)*	(4844)*	(0.46)	(29.91)	-	(5.00)	-
	[1.00]	[1.00]	[1.00]*	[1.00]*	[1.00]	[1.00]	-	[0.68]	-

Human faces: Our DC-decomposition finds the global optimum in 93.59% of the images. Figure 2 shows the three images in which the relative energy difference is largest. Surprisingly the global optimum of the energy function does not match the reality in this images in contrast to DC_{MAX}.

Human bodies: Detecting human bodies is a very challenging task, see Figure 3. While in the second image the solution, achieved by DC_{MAX}, does not describe the underlying image accurately, in the third image it is closer to the human perception. For this complex models DC_{MAX} ends up in 39.26% of the images in the global optimum.

5 Conclusions

We introduced a novel class of approximate MRF-inference algorithms based on quadratic DC-programming. Besides provable convergence properties, the approach shows competitive performance. It is applicable to highly-connected

Table 4. shows the mean energy, time in seconds, and accuracy with respect to the global optimum. For real world examples DC_{MAX} outperforms TRBP, but finds for human bodies configurations with slightly higher energies than BP.

Experiment	A^*	DC_{MAX}	BP	TRBP
Face	45.7410	45.7669	46.7476	46.8325
	(0.0003)	(3.7384)	(0.0070)	(0.0074)
	[1.0000]	[0.9538]	[0.9701]	[0.7658]
Body	57.6411	60.5376	58.2711	73.8115
	(5.6569)	(79.0601)	(0.3382)	(1.4921)
	[1.0000]	[0.6010]	[0.8673]	[0.4057]

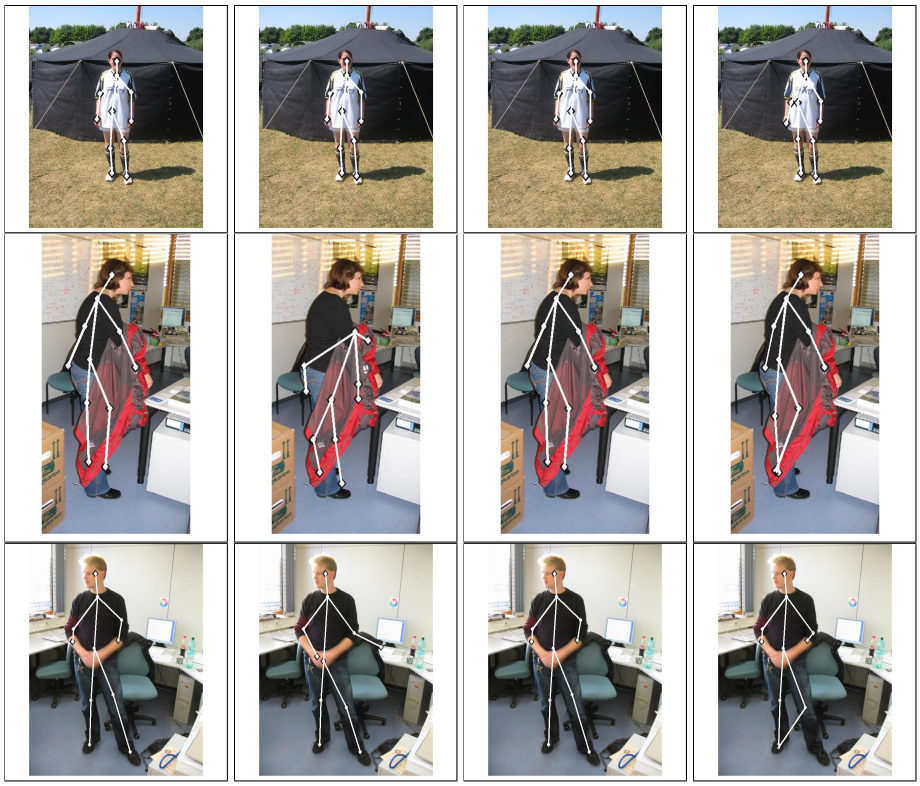

Fig. 3. The pictures above show 3 typical images. In the first one DC_{MAX} found the global optima, in the second it stopped in a critical point which describes the image not well, in the third the solution of DC_{MAX} describes the image better than the global optimum. From left the configurations of A^*, DC_{MAX}, BP and TRBP are shown.

graphical models where standard LP-based relaxations cannot be applied, because the number of constraints becomes too large.

Our future work will supplement the DC-programming framework by globalization strategies and focus on the derivation of performance bounds that hold for any application.

References

1. Mosek 5.0 - http://www.mosek.com.
2. Le Thi Hoai An and Pham Dinh Tao. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133:23–46(24), January 2005.
3. M. Bergtholdt, J.H. Kappes, and C. Schnörr. Learning of graphical models and efficient inference for object class recognition. In *28th Annual Symposium of the German Association for Pattern Recognition*, September 2006.
4. R. Horst and N.V. Thoai. DC programming: Overview. *J. Optimiz. Theory Appl.*, 103(1):1–43, 1999.
5. V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Patt. Anal. Mach. Intell.*, 28(10):1568–1583, 2006.
6. V. Kolmogorov and C. Rother. Comparison of energy minimization algorithms for highly connected graphs. In *Proc. ECCV*, 2006.
7. N. Komodakis and G. Tziritas. Approximate labeling via graph cuts based on linear programming. *PAMI*, 29(8):2649–2661, 2007.
8. M. P. Kumar, V. Kolmogorov, and P. H. S. Torr. An analysis of convex relaxations for MAP estimation. In *Proceedings of Advances in Neural Information Processing Systems*, 2007.
9. M. Pawan Kumar, P. H. S. Torr, and A. Zisserman. Solving markov random fields using second order cone programming relaxations. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1045–1052, Washington, DC, USA, 2006. IEEE Computer Society.
10. Pradeep Ravikumar and John Lafferty. Quadratic programming relaxations for metric labeling and markov random field MAP estimation. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 737–744, New York, NY, USA, 2006. ACM Press.
11. 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. In *Proc. ECCV*, 2006.
12. M.J. Wainwright, T.S. Jaakola, and A.S. Willsky. MAP estimation via agreement on trees: message-passing and linear programming. *IEEE Trans. Inform. Theory*, 51(11):3697– 3717, 2005.
13. T. Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Patt. Anal. Mach. Intell.*, 29(7):1165–1179, 2007.
14. Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005.
15. A. L. Yuille. CCCP algorithms to minimize the bethe and kikuchi free energies: convergent alternatives to belief propagation. *Neural Comput.*, 14(7):1691–1722, 2002.