A memory gradient algorithm for l2-l0 regularization with applications to image restoration

Emilie Chouzenoux, Jean-Christophe Pesquet, Hugues Talbot, Anna Jezierska

A MEMORY GRADIENT ALGORITHM FOR $\ell_2 - \ell_0$ REGULARIZATION WITH APPLICATIONS TO IMAGE RESTORATION

Emilie Chouzenoux, Jean-Christophe Pesquet, Hugues Talbot and Anna Jezierska

Université Paris-Est, Lab. d'Informatique Gaspard Monge, UMR CNRS 8049
77454 Marne la Vallée Cedex 2, France

ABSTRACT

In this paper, we consider a class of differentiable criteria for sparse image recovery problems. The regularization is applied to a linear transform of the target image. As special cases, it includes edge preserving measures or frame analysis potentials. As shown by our asymptotic results, the considered $\ell_2 - \ell_0$ penalties may be employed to approximate solutions to ℓ_0 penalized optimization problems. One of the advantages of the approach is that it allows us to derive an efficient Majorize-Minimize Memory Gradient algorithm. The fast convergence properties of the proposed optimization algorithm are illustrated through image restoration examples.

Index Terms— Non-convex optimization, edge preservation, sparse representations, Majorize-Minimize algorithms, inverse problems, denoising, deblurring.

1. INTRODUCTION

In this work, we consider a wide range of inverse problems where an image $\hat{\boldsymbol{x}} \in \mathbb{R}^N$ can be efficiently estimated from degraded data $\boldsymbol{y} \in \mathbb{R}^Q$ by using a class of sparsity promoting regularized criteria. In many computed imaging applications, the observations are related to the original image $\overline{\boldsymbol{x}}$ through a linear model of the form $\boldsymbol{y} = \boldsymbol{H}\overline{\boldsymbol{x}} + \boldsymbol{u}$, where $\boldsymbol{H}$ is a matrix in $\mathbb{R}^{Q\times N}$ that models the degradation process (e.g., a convolution operator) and $\boldsymbol{u}$ is an additive noise vector. A usual approach is to recover $\overline{\boldsymbol{x}}$ by solving the following penalized optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \big(F_\delta(\boldsymbol{x}) = \Phi(\boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}) + \Psi_\delta(\boldsymbol{x})\big), \tag{1}$$

which combines a data-fidelity term Φ and a regularization term Ψ_δ parameterized by a constant $\delta > 0$. In this work, we will be interested in the case when Φ is a function with a Lipschitz continuous gradient. An important particular case is when Φ is the squared Euclidean norm. The problem then reduces to a penalized least squares (PLS) problem [1]. Another case of interest is when Φ is the separable Huber function [2, Example 5.4] which is useful to limit the influence of outliers in the observed data.

An efficient strategy to promote images formed by smooth regions separated by sharp edges, is to use regularization terms of the form

$$\Psi_\delta(\boldsymbol{x}) = \lambda \sum_{c=1}^{C} \psi_\delta(\boldsymbol{V}_c^\top \boldsymbol{x}) + \|\boldsymbol{\Pi}\boldsymbol{x}\|^2, \tag{2}$$

where $\mathcal{V} = \big\{\boldsymbol{V}_c, c \in \{1,\ldots,C\}\big\}$ is a dictionary of analysis vectors in $\mathbb{R}^N$, $\lambda > 0$ is a weighting factor and $\boldsymbol{\Pi}$ is a matrix in $\mathbb{R}^{P\times N}$. For edge preserving purposes, $\boldsymbol{V}_c$ may be a vector serving to compute a discrete difference between neighboring pixels. The final quadratic penalty term in (2) may play a role similar to the elastic net regularization introduced in [3]. We seek to establish some of the theoretical properties of this framework.

In order to preserve significant coefficients in $\mathcal{V}$, ψ_δ should have a slower than parabolic growth, as this reduces the cost associated with these components. Two of the main families of such functions known in the literature are:

- $\ell_2 - \ell_1$ functions, i.e. convex, continuously differentiable, asymptotically linear functions with a quadratic behavior near 0. A typical example is the hyperbolic function $(\forall t \in \mathbb{R})\ \psi_\delta(t) = \sqrt{t^2 + \delta^2}$, $\delta > 0$. In the limit case when $\delta \to 0$, the classical ℓ_1 norm is obtained.
- $\ell_2 - \ell_0$ functions, i.e. asymptotically constant functions with a quadratic behavior near 0. A typical example is the truncated quadratic function $(\forall t \in \mathbb{R})\ \psi_\delta(t) = \delta^{-2}\min(t^2, \delta^2)$, $\delta > 0$. When $\delta \to 0$, the ℓ_0 penalty is obtained.

The $\ell_2 - \ell_0$ approach is popular in the literature [4, 5, 6], due to its ability to better preserve edges between homogeneous regions. However, the non-convexity and sometimes non-differentiability of the potential function lead to a difficult optimization problem.

Here, we will consider a class of differentiable potential functions, which can be viewed as smoothed versions of the truncated quadratic penalization. The rest of the paper is organized as follows: properties of the considered optimization

This work was supported by the Agence Nationale de la Recherche under grant ANR-09-EMER-004-03.

problem are first investigated in Section 2. Then, we introduce in Section 3 an original minimization strategy based on a memory gradient scheme. We also discuss the general convergence properties of our algorithm. Finally, Section 4 illustrates the good performance of the algorithm through a set of comparisons and experiments in image restoration.

2. CONSIDERED CLASS OF CRITERIA

We will focus on potentials satisfying the following properties:

Assumption 1.

(i) $(\forall \delta \in (0,+\infty))$ ψ_δ *is differentiable.*

(ii) $(\forall (\delta_1,\delta_2) \in (0,+\infty)^2)$ $\delta_1 \leq \delta_2 \Rightarrow (\forall t \in \mathbb{R})$ $\psi_{\delta_1}(t) \geq \psi_{\delta_2}(t) \geq 0$.

(iii) $(\forall t \in \mathbb{R})$ $\lim_{\delta \to 0} \psi_\delta(t) = \begin{cases} 0 & \text{if } t = 0 \\ 1 & \text{otherwise.} \end{cases}$

The latter property shows that the ℓ_0 penalty is asymptotically obtained. Examples of functions satisfying the above assumptions are the following

- $(\forall t \in \mathbb{R})$ $\psi_\delta^{(1)}(t) = \frac{t^2}{2\delta^2+t^2}$
- $(\forall t \in \mathbb{R})$ $\psi_\delta^{(2)}(t) = 1 - \exp(-\frac{t^2}{2\delta^2})$

which can be viewed as smoothed versions of the truncated quadratic function (see Fig. 1).

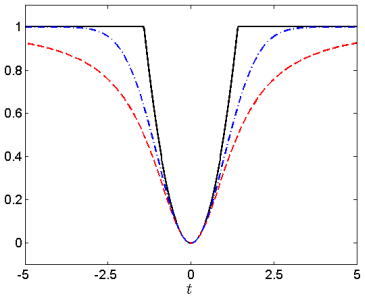

Fig. 1. Truncated quadratic penalty (black, full) and its smoothed approximations $\psi_\delta^{(1)}$ (red, dashed) and $\psi_\delta^{(2)}$ (blue, dash-dot).

We have then the following result:

Proposition 1. *Assume that* Φ *in* (1) *is coercive (that is* $\lim_{\|x\| \to +\infty} \Phi(x) = +\infty$*) and that* $\operatorname{Ker} \boldsymbol{H} \cap \operatorname{Ker} \boldsymbol{\Pi} = \{\boldsymbol{0}\}$*, where* $\operatorname{Ker} \boldsymbol{H}$ *and* $\operatorname{Ker} \boldsymbol{\Pi}$ *are the nullspaces of* $\boldsymbol{H}$ *and* $\boldsymbol{\Pi}$*, respectively. Then, for every* $\delta > 0$*,* F_δ *has a minimizer.*

Note that, in the particular case when $\boldsymbol{H}$ is injective (e.g. in denoising applications), the existence of a minimizer is ensured if $\boldsymbol{\Pi} = \boldsymbol{0}$. Asymptotic convergence can also be obtained by using epi-convergence of F_δ to the following ℓ_0-penalized objective function:

$$F_0 \colon \boldsymbol{x} \mapsto \Phi(\boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}) + \lambda \ell_0(\boldsymbol{V}\boldsymbol{x}) + \|\boldsymbol{\Pi}\boldsymbol{x}\|^2, \qquad (3)$$

where $\boldsymbol{V}^\top = [\boldsymbol{V}_1 \,|\, \ldots \,|\, \boldsymbol{V}_C]$.

Proposition 2. *Let* $(\delta_n)_{n\in\mathbb{N}}$ *be a decreasing sequence of positive real numbers converging to* 0*. Under the same assumptions as in Proposition 1,* $\inf F_{\delta_n} \to \inf F_0$ *as* $n \to +\infty$*. In addition, if for every* $n \in \mathbb{N}$ $\widehat{\boldsymbol{x}}_n$ *is a minimizer of* F_{δ_n}*, then the sequence* $(\widehat{\boldsymbol{x}}_n)_{n\in\mathbb{N}}$ *is bounded and all its cluster points are minimizers of* F_0*.*

The above proposition justifies theoretically that a minimizer of F_0 can be well-approximated by choosing a small enough δ.

3. PROPOSED OPTIMIZATION ALGORITHM

3.1. Memory gradient algorithm

A classical optimization strategy consists of building a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ of $\mathbb{R}^N$ in order to iteratively decrease the criterion F_δ. This can be performed by moving the current solution $\boldsymbol{x}_k$ at iteration $k \in \mathbb{N}$ along a direction $\boldsymbol{d}_k \in \mathbb{R}^N$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k, \qquad (4)$$

where $\alpha_k > 0$ is the *stepsize* and $\boldsymbol{d}_k$ is a *descent direction* i.e., it satisfies $\boldsymbol{g}_k^\top \boldsymbol{d}_k < 0$ where $\boldsymbol{g}_k$ denotes the gradient of F_δ at $\boldsymbol{x}_k$.

The simplest choice for the direction is the opposite of the gradient of the criterion at the current point, which leads to the steepest descent algorithm. A significant improvement of the convergence rate is achieved by the nonlinear conjugate gradient (NLCG) algorithm, where the direction results from a linear combination of the opposite gradient with the previous direction:

$$\boldsymbol{d}_k = -\boldsymbol{g}_k + \beta_k \boldsymbol{d}_{k-1}. \qquad (5)$$

One of the central point in the development of NLCG methods is to define a suitable value of β_k based on certain conjugacy principles [7]. In the memory gradient (MG) method introduced in [8], the optimization scheme (4)-(5) is reformulated as:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{s}_k, \qquad (6)$$

where $\boldsymbol{D}_k = [-\boldsymbol{g}_k \quad \boldsymbol{x}_k - \boldsymbol{x}_{k-1}] \in \mathbb{R}^{N\times 2}$ denotes a set of search directions and $\boldsymbol{s}_k \in \mathbb{R}^2$ a multivariate stepsize that aims at partially minimizing $f_\delta \colon \boldsymbol{s} \mapsto F_\delta(\boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{s})$.

Recently, the MG scheme (6) has been shown to outperform standard descent algorithms such as NLCG over a set of PLS minimization problems [9, 10]. The convergence of the recursive update equation (6) requires the design of a proper strategy to determine the stepsize $\boldsymbol{s}_k$, which we discuss in the next section.

3.2. Majorize-Minimize stepsize

The minimization of f_δ using the Majorization-Minimization (MM) principle is perfomed by successive minimizations of

tangent majorant functions for f_δ. A function $q(., \boldsymbol{s}')$ is said to be a tangent majorant for f_δ at $\boldsymbol{s}'$ if for all $\boldsymbol{s}$,

$$q(\boldsymbol{s}, \boldsymbol{s}') \geqslant f_\delta(\boldsymbol{s}) \quad \text{and} \quad q(\boldsymbol{s}', \boldsymbol{s}') = f_\delta(\boldsymbol{s}'). \tag{7}$$

Following [9], we propose the quadratic tangent majorant function of the form:

$$q(\boldsymbol{s}, \boldsymbol{s}') = f_\delta(\boldsymbol{s}') + \nabla f_\delta(\boldsymbol{s}')^\top (\boldsymbol{s} - \boldsymbol{s}') + \frac{1}{2} (\boldsymbol{s} - \boldsymbol{s}')^\top \boldsymbol{B}_{\boldsymbol{s}'} (\boldsymbol{s} - \boldsymbol{s}'), \tag{8}$$

where $\boldsymbol{B}_{\boldsymbol{s}'}$ is a 2×2 symmetric positive definite matrix that ensures the fulfillment of majorization properties (7). The initial minimization of f_δ is replaced by a sequence of easier subproblems, corresponding to the MM update rule:

$$\begin{cases} \boldsymbol{s}_k^0 = \boldsymbol{0}, \\ \boldsymbol{s}_k^j = \arg\min_{\boldsymbol{s}} q(\boldsymbol{s}, \boldsymbol{s}_k^{j-1}), \quad j \in \{1, \ldots, J\}, \\ \boldsymbol{s}_k = \boldsymbol{s}_k^J. \end{cases} \tag{9}$$

Let us now make the following additional assumptions:

Assumption 2.

(i) *The gradient of Φ is L-Lipschitzian.*[1]

(ii) *ψ_δ is even.*

(iii) *$\psi_\delta(\sqrt{.})$ is concave on $\mathbb{R}^+$.*

(iv) *There exists $\overline{\omega} \in [0, +\infty)$ such that $(\forall t \in (0, +\infty))$ $0 \leq \dot{\psi}_\delta(t) \leq \overline{\omega} t$ where $\dot{\psi}_\delta$ is the derivative of ψ_δ. In addition, $\lim_{\substack{t \to 0 \\ t \neq 0}} \dot{\psi}_\delta(t)/t \in \mathbb{R}$.*

We emphasize the fact that Assumptions 2(ii)-(iv) hold for the ℓ_2-ℓ_0 penalties $\psi_\delta^{(1)}$ and $\psi_\delta^{(2)}$. Let us also introduce

$$\boldsymbol{A}(\boldsymbol{x}) = \mu \boldsymbol{H}^\top \boldsymbol{H} + 2 \boldsymbol{\Pi}^\top \boldsymbol{\Pi} + \lambda \boldsymbol{V}^\top \mathrm{Diag}\{\boldsymbol{b}(\boldsymbol{x})\} \boldsymbol{V}, \tag{10}$$

where $\mu \in (L, +\infty)$ and $\boldsymbol{b}(\boldsymbol{x})$ is a C dimensional vector with entries: $(\forall c \in \{1, \ldots, C\})$ $b_c(\boldsymbol{x}) = \omega_\delta(\boldsymbol{V}_c^\top \boldsymbol{x})$, where $(\forall t \in \mathbb{R})$ $\omega_\delta(t) = \dot{\psi}_\delta(t)/t$ (the function ω_δ is extended by continuity at 0). According to [11], taking

$$\boldsymbol{B}_{\boldsymbol{s}_k^j} = \boldsymbol{D}_k^\top \boldsymbol{A}(\boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{s}_k^j) \boldsymbol{D}_k$$

ensures that $q(., \boldsymbol{s}_k^j)$ is a tangent majorant for f_δ at $\boldsymbol{s}_k^j$. Hence, given (8), we obtain an explicit stepsize formula:

$$\boldsymbol{s}_k^{j+1} = \boldsymbol{s}_k^j - \boldsymbol{B}_{\boldsymbol{s}_k^j}^{-1} \nabla f_\delta(\boldsymbol{s}_k^j).$$

[1] This means that $(\forall \boldsymbol{x} \in \mathbb{R}^N)(\forall \boldsymbol{y} \in \mathbb{R}^N)$ $\|\nabla \Phi(\boldsymbol{x}) - \nabla \Phi(\boldsymbol{y})\| \leqslant L \|\boldsymbol{x} - \boldsymbol{y}\|$.

3.3. Convergence result

Proposition 3. *Under the same assumptions as in Proposition 1, for all $J \geqslant 1$, the MM-MG algorithm given by* (6) *and* (9) *converges in the sense that $\lim_{k \to \infty} \boldsymbol{g}_k = \boldsymbol{0}$. Furthermore, under some technical assumptions, it can be proved that the sequence $(\boldsymbol{x}_k)_{k \in \mathbb{N}}$ converges to a critical point $\widetilde{\boldsymbol{x}}$ of F_δ.*

Note that the computation of the MM stepsize requires specifying the number of MM sub-iterations J. In our experiments, $J = 1$ was observed to yield the best results in terms of convergence profile.

4. APPLICATION TO EDGE-PRESERVING IMAGE RESTORATION

Two image restoration scenarios are considered. In the first one, $\boldsymbol{y}$ is a noisy, blurred image generated by using a Gaussian point spread function of standard deviation 2.24 and of size 17×17, with the convolution product implemented using the Dirichlet boundary condition. In the second case, only noise is added to the original image. In both cases, the noise is white additive Gaussian with standard deviation σ_u, and the original image is the standard `Elaine` image of size $N = 512 \times 512$. An analysis-based PLS criterion is considered by taking $\Phi = \|\cdot\|^2$ and $\boldsymbol{\Pi} = \tau \mathbf{I}$ with $\tau \in \mathbb{R}$ in (1) and (2). In our experiments, $\boldsymbol{V}$ is simply the first-order difference matrix (i.e. the discrete gradient computed in the horizontal and vertical directions), $\psi_\delta = \psi_\delta^{(2)}$, and $\boldsymbol{H}$ is the blur operator or the identity matrix respectively. This criterion depends on the parameters λ, δ and τ. With deblurring applications, the convolution matrix $\boldsymbol{H}$ is not necessarily injective. Thus, we set τ equal to a small positive value in order to fulfill the assumptions of Propositions 1-3. In the denoising case, τ is set to zero. Parameters λ and δ are tuned to maximize the SNR between the original image $\overline{\boldsymbol{x}}$ and its reconstructed version $\widehat{\boldsymbol{x}}$. In Figs. 2 and 3, the reconstructed images and their SNR, MSSIM [12] are displayed. They were obtained with the MM-MG algorithm initialized with a uniform zero image.

Table 1 shows the computational times achieved with several optimization strategies applied to the denoising problem. These results were obtained with C codes running on a single-core Intel Xeon 2.5GHz with 32GB of RAM (RedHat Enterprise Linux 5.5). First, we compare the MM-MG algorithm given by (6) and (9) where $J = 1$, with the Beck-Teboulle (BT) gradient-based algorithm [13] and with the fast version of half quadratic (HQ) algorithm [11]. In the latter, the inner optimization problems are solved partially with a conjugate gradient algorithm. For each algorithm, the global stopping rule is $\|\boldsymbol{g}_k\|/\sqrt{N} < 10^{-4}$. For this setting, no significant differences between algorithms have been observed in terms of reconstruction quality. According to Table 1, our method outperforms both BT and HQ algorithms in terms of convergence speed. Similar conclusions can be drawn for the deconvolution problem (MM-MG: 11 s, BT: 30 s, HQ: 35 s).

Fig. 2. Noisy image with SNR=15 dB, MSSIM = 0.59, $\sigma_u = 25.2$ (left) and denoised image (right) with MM-MG algorithm using $\lambda = 2028$, $\delta = 30$, SNR=24.4 dB, MSSIM = 0.87.

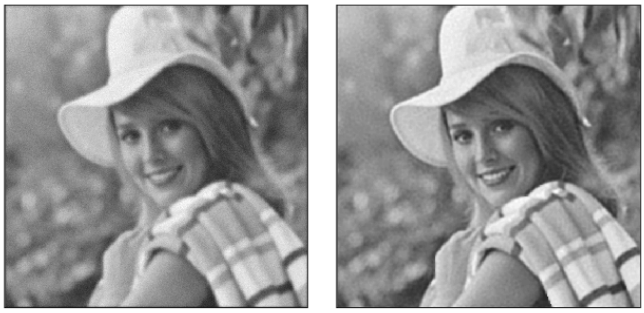

Fig. 3. Noisy blurred image with SNR=21.8 dB, MSSIM = 0.88, $\sigma_u = 4.4$ (left) and deblurred image (right) with MM-MG algorithm using $\lambda = 200$, $\delta = 50$, $\tau = 10^{-10}$, SNR=25.5 dB, MSSIM = 0.92.

Proposed MM-MG algorithm	28 s
Beck-Teboulle algorithm [13]	45 s
Half-Quadratic algorithm [11]	97 s
Tree-Reweighted algorithm [14]	181 s
Belief Propagation algorithm [15]	1958 s

Table 1. Convergence speed of several optimization algorithms for the considered denoising problem.

We also consider the minimization of (1) for the denoising case only, using two state-of-the-art combinatorial optimization algorithms, when ψ_δ is the truncated quadratic penalty. Both considered algorithms lead to a SNR$= 24.3$ dB for the recovered image, which is very similar to the one obtained with smooth regularization. However, Table 1 shows that they are more demanding in terms of computational time than MM-MG. Moreover, to our knowledge, versions of the combinatorial algorithms which would be applicable to the deblurring problem are not available.

5. CONCLUSION

In this work, we have considered image restoration problems with non-convex $\ell_2 - \ell_0$ regularization terms. These penalties are similar in principle to truncated quadratic terms that are widely used in combinatorial optimization methods, but are regular enough to allow us to propose an efficient and effective Majorization-Minimization Memory Gradient algorithm. In an image denoising application, we showed that our approach obtains almost identical results as state-of-the-art optimization strategies but outperforms them significantly in term of speed. In addition, our framework appears also to be quite competitive for deblurring.

6. REFERENCES

[1] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Trans. Image Processing*, vol. 6, pp. 298–311, 1997.

[2] P. J. Huber, *Robust Statistics*, John Wiley, New York, NY, 1981.

[3] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Statist. Soc. B*, vol. 67, no. 2, pp. 301–320, 2005.

[4] A.H. Delaney and Y. Bresler, "Globally convergent edge-preserving regularized reconstruction: an application to limited-angle tomography," *IEEE Trans. Image Processing*, vol. 7, no. 2, pp. 204 –221, February 1998.

[5] O. Veksler, "Graph cut based optimization for MRFs with truncated convex priors," in *IEEE Conf. Comp. Vision Pattern Recogn.*, Minneapolis, 17-22 June 2007, pp. 1–8.

[6] Y. Zhang and N. Kingsbury, "Restoration of images and 3D data to higher resolution by deconvolution with sparsity regularization," in *IEEE Conf. on Image Processing*, Hong Kong, 26-29 September 2010, pp. 1685 –1688.

[7] W. W. Hager and H. Zhang, "A survey of nonlinear conjugate gradient methods," *Pacific J. Optim.*, vol. 2, no. 1, pp. 35–58, January 2006.

[8] A. Miele and J. W. Cantrell, "Study on a memory gradient method for the minimization of functions," *J. Optim. Theory Appl.*, vol. 3, no. 6, pp. 459–470, 1969.

[9] E. Chouzenoux, J. Idier, and S. Moussaoui, "A Majorize-Minimize strategy for subspace optimization applied to image restoration," *IEEE Trans. Image Processing*, 2010, to appear.

[10] M. Zibulevsky and M. Elad, "$\ell_2 - \ell_1$ optimization in signal and image processing," *IEEE Signal Processing Mag.*, vol. 27, no. 3, pp. 76–88, May 2010.

[11] M. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms," *IEEE Trans. Image Processing*, vol. 15, no. 5, pp. 1130–1142, 2006.

[12] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600 –612, April 2004.

[13] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Processing*, vol. 18, no. 11, pp. 2419 –2434, November 2009.

[14] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1568 –1583, October 2006.

[15] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Computer Vision*, vol. 70, pp. 41–54, 2006.