Convex Multi-Class Image Labeling by Simplex-Constrained Total Variation

Jan Lellmann, Jörg Kappes, Jing Yuan, Florian Becker, and Christoph Schnörr

Image and Pattern Analysis Group (IPA)
HCI, Dept. of Mathematics and Computer Science, University of Heidelberg
{lellmann,kappes,yuanjing,becker,schnoerr}@math.uni-heidelberg.de

Abstract. Multi-class labeling is one of the core problems in image analysis. We show how this combinatorial problem can be approximately solved using tools from convex optimization. We suggest a novel functional based on a multidimensional total variation formulation, allowing for a broad range of data terms. Optimization is carried out in the operator splitting framework using Douglas-Rachford Splitting. In this connection, we compare two methods to solve the Rudin-Osher-Fatemi type subproblems and demonstrate the performance of our approach on single- and multichannel images.

1 Introduction

In this paper, we study the variational approach

$$\inf_{u \in C} f(u) \;, \quad f(u) = -\int_\Omega \langle u(x), s(x) \rangle dx + \lambda \, \mathrm{TV}(u) \;, \quad \lambda > 0 \;, \tag{1}$$

for determining a labeling $u : \Omega \to \mathbb{R}^L$, that is a contextual classification of each pixel $x \in \Omega$ into one out of L classes, based on an arbitrary vector-valued similarity function $s(x) \in \mathbb{R}^L$ as input data that has been computed from image data beforehand.

The objective function (1) comprises the common form of a data term plus a regularization term. The data term is given by the L^2 inner product of the assignment variables u and the similarity function s, and the regularizer is a total variation (TV) formulation for vector-valued data,

$$\mathrm{TV}(u) = \int_\Omega \sqrt{\|\nabla u_1\|^2 + \cdots + \|\nabla u_L\|^2} dx \;. \tag{2}$$

Furthermore, the constraint $u \in C$ restricts the vector field $u(x)$ at each location $x \in \Omega$ to lie in the standard probability simplex, that is $u(x) \in \mathbb{R}^L_+$ and $\sum_{i=1}^L \big(u(x)\big)_i = 1$ for all $x \in \Omega$.

Our work is motivated by the following observation. Suppose that at each pixel $x \in \Omega$, there is an *unambiguous* assignment (labeling) of the data $s(x)$ to some class $l \in \{1, \ldots, L\}$ represented by the corresponding l-th unit vector,

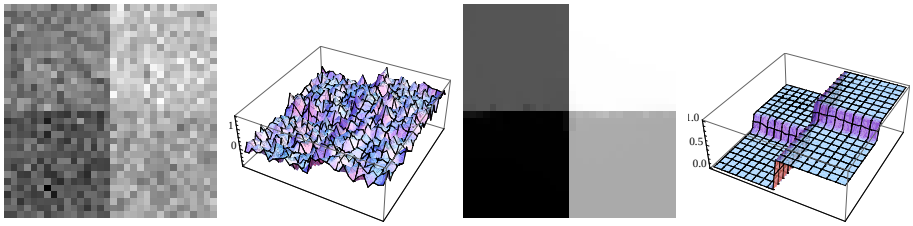

Fig. 1. Left: Noisy input image. **Right:** The labeled image based on the non-binary assignment u as global minimizer of the convex approach (1). The discrete problem is accurately solved by a continuous approach.

$u(x) = e^l$. Then, an interface with area A between two image regions labeled with l and l', respectively, adds $A\sqrt{2}$ to the regularization term iff $l \neq l'$, as all but two gradients under the square root vanish. As a result, under these assumptions and up to the immaterial constant $\sqrt{2}$, the TV term corresponds to the well-known Potts model that assigns constant penalties to local changes of the labeling.

A significant difference between the Potts model and our approach (1), however, is that the former amounts to solve a *discrete combinatorial* problem, whereas the latter is a *continuous convex* optimization problem. Experiments show that our approach (1) approximates *discrete* decisions fairly well (Fig. 1 and 2) by computing a global optimum to a single convex optimization problem. By contrast, the state-of-the-art discrete approach [1] approximates the combinatorial solution by solving a non-uniquely defined *sequence* of binary problems via graph cuts. This fact, along with the potential of continuous convex optimization for parallel implementations and their more robust dependency on (hyper-) parameters, motivated to investigate the approach (1) as a promising model for a general "labeling submodule" within computer vision systems. To this end,

- We have a closer look at the data and regularization terms (section 2).
- We apply an operator splitting approach to (1) in order to decompose the computation of a globally optimal labeling into two independent computational steps: TV denoising for vector-valued data, and projection of the labeling vectors $u(x)$ on the canonical simplex (section 3).
- We evaluate two different algorithms for the TV denoising subroutine (section 4) and compare the performance of our convex method to a range of established graph cut-based approaches (section 5).

Related work. In contrast to the binary case with anisotropic discretization [2], multi-class energies are generally not submodular and thus cannot be optimized globally using graph cuts [3]. Some extensions exist, which find a local minimum by solving a sequence of binary graph cuts [1]. The continuous formulation – optimization on the set of characteristic functions – is known as *continuous cut* [5]. Chan et al. [6] showed that this problem can be relaxed and solved on a convex set, without losing global optimality. In contrast, our work is aimed at

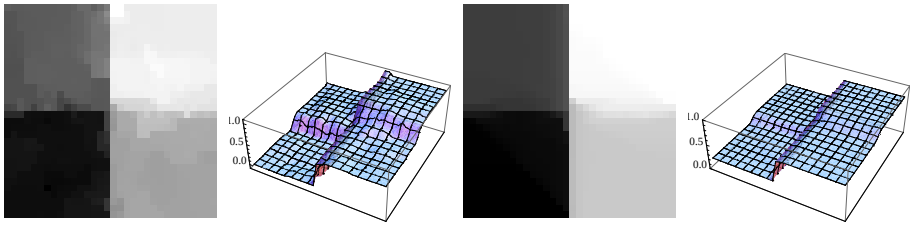

Fig. 2. Output of the standard TV approach [4] for scalar-valued images applied to the noisy input image depicted in Fig. 1, for different values of the regularization parameter λ. Irrespective of this value, the performance is worse than with the approach (1) (cf. Fig. 1, right), because the latter approximates the Potts model that does *not* depend on the size (contrast) of discontinuities. Consequently, the former approach cannot remove noise without degrading weak discontinuities, as is apparent above for the horizontal discontinuities.

the multi-class case. In [7], a comparable approach based on [8] was presented, which relies on a natural ordering of the labels, as given in e.g. stereo reconstruction. An approach very similar to ours was recently presented in [9], where the authors use a different formulation of the total variation on vector fields, and an alternating optimization method. The (discrete) Potts model was studied in [10], where approximate solutions were computed by an LP relaxation with explicit constraints. In contrast, our approach considers the general TV term and a problem decomposition into efficiently solvable subproblems, without the need to introduce additional variables.

Notation. We consider the *discretized* version of our approach (1). Let $\Omega = \{1,\dots,n_1\} \times \cdots \times \{1,\dots,n_d\} \subseteq \mathbb{R}^d, d \in \mathbb{N}$, denote a regular image grid of $n := |\Omega|$ pixels. The (multidimensional) image space $X := \mathbb{R}^{n\times L}$ is equipped with the Euclidean inner product $\langle\cdot,\cdot\rangle_\Omega$ over the vectorized elements. We naturally identify $v = (v^1,\dots,v^L) \in \mathbb{R}^{n\times L}$ with $((v^1)^\top \cdots (v^L)^\top)^\top \in \mathbb{R}^{nL}$. Superscripts v^i denote a collection of vectors, while subscripts v_k denote vector components. Using the notation $e = (1,1,\dots,1)^\top$, the standard simplex on $\mathbb{R}^L$ and its extension C on $\mathbb{R}^{n\times L}$ are given by $\Delta_L := \left\{v \in \mathbb{R}^L \,\middle|\, v \geq 0\,,\ \langle e, v\rangle = 1\right\}$ and $C := \prod_{x\in\Omega} \Delta_L$. Define $\delta_C(x)$ to be 0 iff $x \in C$, and $+\infty$ otherwise.

Let $\mathrm{grad} := (\mathrm{grad}_1^\top,\dots,\mathrm{grad}_d^\top)^\top$ be the d-dimensional forward difference gradient operator for Neumann boundary conditions. Accordingly, $\mathrm{div} := -\mathrm{grad}^\top$ is the backward difference divergence operator for Dirichlet boundary conditions. These operators extend to $\mathbb{R}^{n\times L}$ via $\mathrm{Grad} := (I_L \otimes \mathrm{grad})$, $\mathrm{Div} := (I_L \otimes \mathrm{div})$, where I_L is the $L \times L$ identity matrix. We will also need the convex sets

$$B_\lambda := \left\{(p^1,\dots,p^L) \in \mathbb{R}^{d\times L} \;\middle|\; \Big(\sum_{i=1}^{L} \|p^i\|_2^2\Big)^{\frac{1}{2}} \leqslant \lambda\right\}, \tag{3}$$

$$D_\lambda := \prod_{x\in\Omega} B_\lambda \subseteq \mathbb{R}^{n\times d\times L}\,,\ E_\lambda := \{u \in \mathbb{R}^{n\times L} | u = \mathrm{Div}\, p\,, p \in D_\lambda\}\,. \tag{4}$$

The discrete total variation on vector-valued data is then defined as

$$\mathrm{TV}(u) := \sigma_{E_1}(u) = \sum_{x\in\Omega} \|G_x u\|_2 \,, \tag{5}$$

where $\sigma_M(u) := \sup_{p\in M}\langle u, p\rangle$ is the support function from convex analysis, and G_x is an $(Ld)\times n$ matrix composed of rows of (Grad) s.t. $G_x u$ gives the gradients of all u_i in x stacked one above the other.

2 Variational Approach

Based on the introduced notation, our novel approach (1) reads

$$\inf_{u\in C} f(u)\,, \quad f(u) = \underbrace{-\langle u, s\rangle_\Omega}_{\text{data term}} + \underbrace{\lambda\,\mathrm{TV}(u)}_{\text{regularization term}}\,, \quad \lambda > 0\,, \tag{6}$$

As the objective function f and the constraint set C are convex, the overall problem is convex as well. We will now define and motivate each term.

Data Term. The data term in (6) is fairly general. Any vector-valued similarity function s can be used, whose components $\big(s(x)\big)_i$ indicate the affinity of some data point at x with class i. As an example, suppose we have image features $g(x)$, $x \in \Omega$, prototypical feature vectors $G = (G^1, \ldots, G^L)$ as well as a distance measure d on the features. We might think of g as a grayscale image, of G as some prototypical gray values, and of d as a quadratic distance measure, possibly derived from a statistical noise model.

The hard assignment of the pixel $x \in \Omega$ to a label (or class) $l(x) \in \{1, \ldots, L\}$ should then be penalized by the distance $d(g(x), G^{l(x)})$ of the corresponding feature to the prototype of the assigned class. Denoting the negative distance by s, and summing up over the image domain, we see that

$$\sum_{x\in\Omega} d\left(g(x), G^{l(x)}\right) = -\sum_{x\in\Omega} \big\langle s(x), u(x)\big\rangle \quad \text{for } u(x) = e^{l(x)}\,. \tag{7}$$

Thus, instead of looking for $l \in \{1, \ldots, L\}^n$, we may equivalently look for $u \in \{e^1, \ldots, e^L\}^n$. However, the right hand side formulation has the advantage that it extends naturally to the *soft* assignment $u \in C$: We may now solve the easier problem of optimizing for u on the *convex* set C.

In our experiments, we chose $d(x, y) = ||x - y||_1$, as the ℓ_1-norm is still convex but known to be more robust against noise and outliers. However, s is not restricted to representing distances. In fact, it may be arbitrarily nonlinear and nonconvex in x and g, and involve nonlocal operations on g. The complexity is completely hidden within the precomputed vector s.

Regularization Term. Recall that the regularizer of (6) is defined (5) as

$$\mathrm{TV}(u) = \sup_{p\in D_1} \langle u, \mathrm{Div}\, p\rangle = \sum_{x\in\Omega} \|G_x u\|_2\,. \tag{8}$$

This definition for vector-valued u parallels the definition of the "isotropic" total variation measure in the scalar-valued case [11, 4, 12]. It is also known as *MTV* [13–15], and was recently studied in [16] in its continuous formulation. Contrary to the anisotropic discretization, where one would substitute the sum of 1-norms in (3), it is less biased towards edges parallel to the axes. See also [17] for an overview of TV-based research and applications.

Optimality. After solving the relaxed problem, it remains to show that a binary solution can be recovered. For the continuous, *binary* case, Chan et al. [6] showed that an exact solution can be obtained by thresholding at almost any threshold. However, their results do not immediately transfer to the discrete *multi*-class case. In particular, the crucial "layer cake" formula holds for ℓ_1-, but not ℓ_2 discretizations of the TV.

Contrary to the binary case, it is not clear which rounding scheme to use for vector-valued u. For our experiments, we chose the final class label for each pixel x as the index l of the maximal $u_l^*(x)$ of the global optimum u^* of (6). This defines a suboptimal discrete solution u_t^*. Bounding the error $f(u_t^*) - f(u_d^*)$ with respect to the unknown discrete optimum u_d^* will be subject of our future work.

3 Optimization

Two basic problems arise concerning the optimization of (6): Nondifferentiability of the objective function due to the TV term, and handling of the simplex constraint $u \in C$. We cope with the latter using the tight *Douglas-Rachford* splitting method as presented in the following section. We refer to [18] for the full derivations.

Douglas-Rachford Splitting. Minimization of a proper, convex, lower-semi-continuous (*lsc*) function $f : X \to \mathbb{R}$ can be regarded as finding a zero of its (necessarily maximal monotone [19, Chap. 12]) subgradient operator $T := \partial f : X \rightrightarrows X$. In the operator splitting framework, ∂f is assumed to be decomposable into the sum of two "simple" operators, $T = A + B$, of which forward and backward steps can practically be computed. Here, we consider the (tight) *Douglas-Rachford-Splitting* iteration [20, 21],

$$z^{k+1} \in (J_{\tau A}(2J_{\tau B} - I) + (I - J_{\tau B}))(z^k)\,, \tag{9}$$

where $J_{\tau T} := (I + \tau T)^{-1}$ is the *resolvent* of T. Under the very general constraint that A and B are maximal monotone and $A + B$ has at least one zero, the sequence (z^k) will converge to a point z, with the additional property that $x := J_{\tau B}(z)$ is a zero of T ([22, Thm. 3.15], [22, Prop. 3.20], [22, Prop. 3.19], [23]).

In particular, for $f = f_1 + f_2$, f_i proper, convex, lsc with $\mathrm{ri}(\mathrm{dom}\, f_1) \cap \mathrm{ri}(\mathrm{dom}\, f_2) \neq \emptyset$ ($\mathrm{ri}(S)$ denoting the relative interior of a set S), it can be shown [19, Cor. 10.9] that $\partial f = \partial f_1 + \partial f_2$, and the ∂f_i are maximal monotone. As $x \in J_{\tau \partial f_i}(y) \Leftrightarrow x = \mathrm{argmin}(2\tau)^{-1}\|x - y\|_2^2 + f_i(x)$, the computation of the resolvents reduces to proximal point optimization problems involving only the f_i.

Application. For our specific problem, we split

$$\inf_{u\in C}(f_1(u)+f_2(u))\,,\ f_1(u)=-\langle u,s\rangle_\Omega+\lambda\mathrm{TV}(u)\,,\ f_2(u)=\delta_C(u)\,. \tag{10}$$

and get the following Douglas-Rachford scheme:

Algorithm 1 Outer loop (Douglas-Rachford)

1: choose some u^0 and a fixed step size $\tau>0$
2: **repeat**
3: solve $u^k \leftarrow \operatorname{argmin}_u\{\frac{1}{2\tau}\|u-z^k\|^2-\langle u,s\rangle+\sigma_{E_\lambda}(u)\}$
4: solve $w^k \leftarrow \operatorname{argmin}_w\{\frac{1}{2\tau}\|w-(2u^k-z^k)\|^2+\delta_C(w)\}$
5: $z^{k+1}\leftarrow z^k+w^k-u^k$
6: **until** $\|u^k-u^{k-1}\|_\infty\leqslant\delta_{\text{outer}}$.

As f is bounded from below on the compact set C and thus attains its minimum. From the remarks in the last section, we get convergence of the scheme for the discrete case: $\delta_C(w)$ and σ_{E_λ} are both proper, convex, lsc with $\operatorname{dom}\sigma_{E_\lambda}=\mathbb{R}^n$ and $\operatorname{ri}(C)\neq\emptyset$. In practice, one has to deal with solutions of the subproblems with limited accuracy. While there are extensions of the convergence result that take these inexact solutions into account [22, Prop. 4.50], they require the subproblems to be solved with increasing accuracy. However, we found that the method generally converged even though these requirements were not met.

4 Inner Loop Optimization

The second subproblem (Alg. 1, step 4) is a projection on the constraint set, $w^k=\Pi_C(2u^k-z^k)$, which requires one projection on the low-dimensional unit simplex Δ_L per $x\in\Omega$. These projections can be computed in a finite number of steps [24]. The first subproblem (step 3) is equivalent to

$$u^k=\operatorname{argmin}_u\frac{1}{2}\|u-(z^k+\tau s)\|^2+(\tau\lambda)TV(u), \tag{11}$$

i.e. an extension to vector vector-valued u of the classical Rudin-Osher-Fatemi (ROF, TV-L^2) problem with regularization parameter $\tau\lambda$. Many methods have been suggested to solve the ROF problem, e.g. PDE, fixpoint, or interior point methods for primal [4, 25], dual [26–28], or mixed [29] formulations.

We evaluate two approaches: First, we will formulate a particularly simple gradient projection method in the operator splitting framework, cf. [30]. This scheme was introduced in [27] and extended to the multidimensional case in [31] (see also [16]). The second approach is based on the fast half-quadratic method of Yang et al. [15].

Forward-backward approach. The optimality condition of step 3, $\tau^{-1}(z^k-u)+s\in\partial\sigma_{E_\lambda}(u)$, can be rewritten as $u=\tau\left(\left(z^k/\tau+s\right)-\Pi_{E_\lambda}\left(z^k/\tau+s\right)\right)$. To compute the projection Π_{E_λ}, we use the dual representation,

$$\Pi_{E_\lambda}(x)=\operatorname*{argmin}_{q\in E_\lambda}\frac{1}{2}\|q-x\|_\Omega^2=\operatorname{Div}\left\{\operatorname*{argmin}_p\frac{1}{2}\|\operatorname{Div}p-x\|_\Omega^2+\delta_{D_\lambda}(p)\right\}\,. \tag{12}$$

Using a simple forward-backward splitting for the inner problem results in the (gradient projection) update rule

$$p^{j+1} = \Pi_{D_\lambda}\left(p - \nu \mathrm{Div}^\top(\mathrm{Div}\, p - x)\right).$$

The projection Π_{D_λ} can be computed explicitly and is separable in x, while the inner part can be computed for all models independently. This opens up the method to parallelization.

Convergence is guaranteed for $\nu < 2/\|\mathrm{Div}^\top \mathrm{Div}\|$ (see e.g. [22, Thm. 3.12]). Extending the argument in [26, Thm. 3.1], we find that $\|\mathrm{div}\| \leqslant \sqrt{4d}$. Accordingly, we may set $\nu < \frac{1}{2d}$. In our experiments, we set $\nu = \frac{0.95}{2d}$ to avoid numerical problems close to the theoretical maximum. Wrapping up, we have

Algorithm 2 Inner loop, forward-backward approach

1: $x \leftarrow \frac{z^k}{\tau} + s$, choose arbitrary $p^0 \in \mathbb{R}^{n\times d\times L}$
2: **repeat**
3: $\quad p^{j+1} = \Pi_{D_\lambda}(p^j - \nu \mathrm{Div}^\top(\mathrm{Div}\, p - x))$
4: **until** $\|p^{j+1} - p^j\|_\infty \leqslant \delta_{\text{inner}}$
5: $u^k \leftarrow \tau(x - \mathrm{Div}\, p^{j+1})$.

Half-quadratic approach. While the forward-backward method is simple and easy to implement, its convergence speed is in practice not satisfactory. As an alternative, we tested an ROF specialization of the general multichannel image restoration method by Yang et al. [15]. Starting from (11), the problem is to find

$$u^k = \mathrm{argmin}_u g(u), \quad g(u) := \frac{\mu}{2}\|u - f\|^2 + TV(u), \tag{13}$$

where $\mu := \frac{1}{\tau\lambda}$ and $f := z^k + \tau s$. Using a half-quadratic approach [32, 33], Yang et al. derive the splitting/penalty formulation

$$(u, y) = \operatorname*{argmin}_{y_x\in\mathbb{R}^{Ld}, x\in\Omega, u\in\mathbb{R}^{nL}} \sum_{x\in\Omega}\left(\|y_x\| + \frac{\beta}{2}\|y_x - G_x u\|^2\right) + \frac{\mu}{2}\|u - f\|_\Omega^2 . \tag{14}$$

The parameter β controls smoothing of the total variation; setting $\beta \geqslant n/(2\varepsilon)$ guarantees ε-suboptimality of the solution of the smoothed problem with respect to the original problem (for a derivation see [18]).

Equation (14) can be solved using alternating minimization w.r.t. u and the auxiliary variables y_x. The latter is highly parallelizable, as it boils down to n separate explicit operations:

$$y_x^{j+1} = \max\left\{\|G_x u\| - \beta^{-1}, 0\right\}(G_x u/\|G_x u\|). \tag{15}$$

On the other hand, minimizing (14) for u amounts to solving

$$\left(\mathrm{Grad}^\top \mathrm{Grad} + (\mu/\beta) I_{(nL)}\right) u^{j+1} = \mathrm{Grad}^\top y^{j+1} + \frac{\mu}{\beta} f, \tag{16}$$

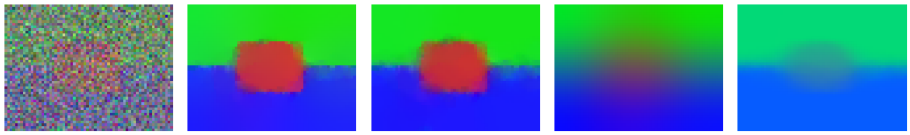

Fig. 3. Results of the speed comparison between forward-backward (FB) and half-quadratic method (HQ) for the inner problem, applied to data from the *first iteration* of the outer problem (cf. Table 1). **Left to right:** Original input, FB with $\tau\lambda = 5$, HQ with $\tau\lambda = 5$, FB with $\tau\lambda = 20$, HQ with $\tau\lambda = 20$. Iteration counts were fixed at 80 resp. 300 to equalize the runtime for both approaches. For larger regularization parameter, the half-quadratic method outperforms the forward-backward approach as smoothness increases.

for u^{j+1}, where y^{j+1} is a proper rearrangement of the y_x.

For periodic boundary conditions, Yang et al. solved (16) rapidly using FFT. In our case, Neumann boundary conditions and thus the Discrete Cosine Transform (DCT-2)[34] are appropriate. This requires $2L$ independent (parallelizable) individual DCTs which can be efficiently computed in $O(n \log n)$ each.

By the alternating application of the above two steps, we can solve (14) for fixed β large enough for any required suboptimality bound. In practice, convergence can be sped up by starting with a small β and solving a sequence of problems for increasing β, warm-starting each with the solution for the previous problem. Given an arbitrary $u^0 \in \mathbb{R}^{nL}$, the complete algorithm reads

Algorithm 3 Inner loop, half-quadratic approach

1: **while** stopping criterium not satisfied **do**
2: compute y^{j+1} from (15)
3: compute u^{j+1} from y^{j+1} and (16),
4: possibly increase β
5: **end while**

The stopping criteria can be based on the residual [15]. For our experiments, we set a fixed iteration count, as increasing β at each step turned out to lead to fastest convergence, and residua for different β are not comparable.

5 Experiments, Performance Evaluation

Inner Problem. We compared the half-quadratic approach to the conventional forward-backward method. The difficulty with the former lies in the choice of the update strategy for β. We chose a generalization of the exponential strategy outlined in [15]: Set $\beta = \beta_{\min}$ and update by multiplying with $c := (\beta_{\max}/\beta_{\min})^{1/K}$ for some K until $\beta = \beta_{\max}$. We made the following observations:

– In order to rapidly minimize the objective function, it is best to use a continuation strategy, i.e. to increase β at each step, rather than spending time on solving (14) exactly for each β.

Table 1. Run times t (in seconds), objective function values r and relative differences $(r_{\mathrm{HQ}} - r_{\mathrm{FB}})/r_{\mathrm{HQ}}$ for the experiment in Fig. 3. For larger $\tau\lambda$, the half-quadratic method gives more accurate results in the same time.

$\tau\lambda$	0.1	1	2	5	10	20	50
t_{HQ}	1.14	1.23	1.20	1.31	0.98	0.95	1.08
t_{FB}	1.03	1.02	1.06	1.03	1.22	1.25	1.19
r_{HQ}	3901.9	27660.7	36778.5	40038.8	42262.8	44377.1	44752.5
r_{FB}	3901.9	27660.4	36760.6	40104.3	42924.3	46988.6	57504.9
rel. diff.	1.17e-16	1.24e-5	4.85e-4	-1.64e-3	-0.0156	-0.0588	-0.285

- Increasing K generally improves the quality of the result.
- For fixed $\beta_{\max}$ and K, there seems to be a unique optimal $\beta_{\min}$ that minimizes the final objective function value.

With the continuation strategy and fixed $\beta_{\max}$, we found the optimal $\beta_{\min}$ to usually lie in the range of $10^{-5}\beta_{\max}$ to $10^{-3}\beta_{\max}$. Unfortunately, there seems to be a strong dependency on the choice of λ as well as the scale and complexity of s. We set $\beta_{\min} = 0.2 \cdot 10^{-4}\beta_{\max}$, which worked well for our data. $\beta_{\max}$ was set at $n/0.2$ according to a suboptimality bound of $\varepsilon = 0.1$ (section 4).

We compared the performance of the two methods in terms of the objective function value for fixed runtime of the optimized Matlab implementations (Fig. 3, Table 1). For larger $\tau\lambda$, the half-quadratic method gives better results. For $\tau\lambda = 20$, less than 10 iterations are required to reach the quality of 300 iterations of the forward-backward method, giving a speedup of about 4-5. However, finding the optimal parameter set is more involved than for the forward-backward method.

Overall Problem. We evaluated the performance of our algorithm against five different methods in their publicly available implementations from the Middlebury MRF benchmark [35]: Belief Propagation (BP), Sequential Belief Propagation (BPS), Graph Cuts with alpha-expansion (GCE) and alpha-beta swap (GCS), and Sequential Tree Reweighted Belief Propagation (TRBPS). Of each of the grayscale 32×32 images, 20 noisy copies were generated and segmented into four gray levels with fixed intensities. In view of the last section and in order not to mix up speed with accuracy issues, we used the forward-backward approach for the inner loop. We set $\delta_{\mathrm{inner}} = 1 \cdot 10^{-3}, \delta_{\mathrm{outer}} = 2 \cdot 10^{-2}$, and $\tau = 1$.

For small λ, our method shows results comparable to the other approaches with respect to the number of bad labels. We point out again that this solution to the non-binary labeling problem is achieved by solving the *convex* optimization problem (6) followed by local rounding as explained in section 2.

In contrast to our method, the MRF benchmark algorithms optimize the *anisotropic* energy. To compensate, their λ was scaled by a common factor of $\approx \sqrt{2}$ that was found empirically. Nevertheless, the discretization gives them a small advantage on images with axis parallel edges (experiments 1 and 2). Figure 6 demonstrates the performance of our algorithm for color segmentation. Only few outer iterations (20 in our case) are necessary for accurate optimization.

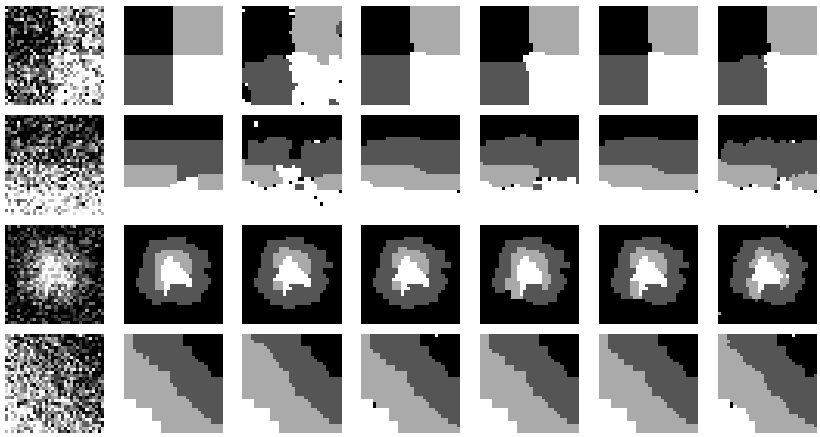

Fig. 4. Exemplary grayscale segmentation results for the benchmarked methods for four labels. **Left to right:** Noisy input data, final results for BP, BPS, GCE, GCS, TRWS, and the proposed method (TV). λ was manually chosen for each method. Axis-parallel edges are better recovered by the anisotropic methods, while our isotropic discretization has an advantage on diagonal edges.

6 Conclusion and Future Work

In this paper, we presented a convex variational approach to solve the combinatorial multi-labeling problem for energies involving a general data term, total-variation-like regularizers, and simplex constraints. To enforce the simplex constraint, we based our approach on the globally convergent Douglas-Rachford operator splitting scheme. We evaluated two methods in order to efficiently solve the ROF-type subproblems, and showed that the half-quadratic approach allows faster convergence at the price of more involved parameter tuning.

Experiments showed that the quality of the generated labelings is comparable to state of the art discrete optimization methods, and can be achieved by just solving a convex optimization problem.

Due to the generality of the data term, our method allows for a wide range of features or distance measures. To fully evaluate these possibilities in connection with variations of the TV measure is a subject of our future research.

Acknowledgements. Jing Yuan gratefully acknowledges support by the German National Science Foundation (DFG) under grant SCHN 457/9-1.

References

1. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI **23**(11) (2001) 1222–1239
2. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. PAMI **26**(9) (2004) 1124–1137

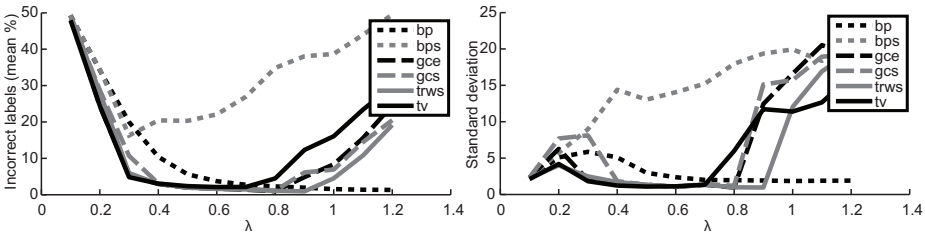

Fig. 5. Error rates for the first experiment in Fig. 4. For each λ, all experiments were repeated 20 times with random noise (zero-mean Gaussian with $\sigma = 0.45, 0.35, 0.25$ resp. 0.35 for experiments 1-4 and image intensities in $[0, 1]$), and the percentage of incorrectly assigned labels compared to ground truth was recorded. Sequential Belief Propagation (BPS) generally performed worst, while our method (TV) was on par with the others, in particular for lower λ. The figure also reveals that belief propagation (BP) gets stuck in a good, but often inferior local optimum, and does not respond to larger values of λ, i.e. stronger regularization requested by the user.

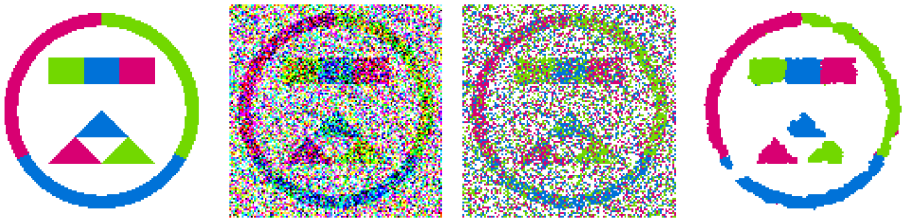

Fig. 6. Performance of our method for four-class segmentation based on ℓ_1 color distance. **Left to right:** Ground truth, inspired by [29, 36]; ground truth overlaid with Gaussian noise, $\sigma = 1$; local nearest-neighbor labeling; our approach with $\lambda = 0.7$ after 20 outer iterations. The energy of the result is about 1% lower than the energy of the ground truth, suggesting that at this noise level, further improvements are limited by the model.

3. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? PAMI **26**(2) (2004) 147–159
4. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60** (1992) 259–268
5. Strang, G.: Maximal flow through a domain. Math. Prog. **26** (1983) 123–143
6. Chan, T.F., Esedoḡlu, S., Nikolova, M.: Algorithms for finding global minimizers of image segmentation and denoising models. J. Appl. Math. **66**(5) (2006) 1632–1648
7. Pock, T., Schönemann, T., Graber, G., Bischof, H., Cremers, D.: A convex formulation of continuous multi-label problems. In: ECCV. Volume 3. (2008) 792–805
8. Ishikawa, H.: Exact optimization for Markov random fields with convex priors. PAMI **25**(10) (2003) 1333–1336
9. Zach, C., Gallup, D., Frahm, J.M., Niethammer, M.: Fast global labeling for real-time stereo using multiple plane sweeps. In: VMV. (2008)
10. Kleinberg, J., Tardos, E.: Approximation algorithms for classification problems with pairwise relationships: Metric labeling and MRFs. In: FOCS. (1999) 14–23

11. Ziemer, W.: Weakly Differentiable Functions. Springer (1989)
12. Meyer, Y.: Oscillating Patterns in Image Processing and Nonlinear Evolution Equations. Volume 22 of Univ. Lect. Series. AMS (2001)
13. Sapiro, G., Ringach, D.L.: Anisotropic diffusion of multi-valued images with applications to color filtering. In: Trans. Image Process. Volume 5. (1996) 1582–1586
14. Chan, T.F., Shen, J.: Image processing and analysis. SIAM (2005)
15. Yang, J., Yin, W., Zhang, Y., Wang, Y.: A fast algorithm for edge-preserving variational multichannel image restoration. Tech. Rep. 08-09, Rice Univ. (2008)
16. Duval, V., Aujol, J.F., Vese, L.: A projected gradient algorithm for color image decomposition. CMLA Preprint (2008-21) (2008)
17. Chan, T., Esedoglu, S., Park, F., Yip, A.: Total variation image restoration: Overview and recent developments. In: The Handbook of Mathematical Models in Computer Vision. Springer (2005)
18. Lellmann, J., Kappes, J., Yuan, J., Becker, F., Schnörr, C.: Convex multi-class image labeling by simplex-constrained total variation. TR, U. of Heidelberg (2008)
19. Rockafellar, R., Wets, R.J.B.: Variational Analysis. 2nd edn. Springer (2004)
20. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. of the AMS **82**(2) (1956) 421–439
21. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM Journal on Numerical Analysis **16**(6) (1979) 964–979
22. Eckstein, J.: Splitting Methods for Monotone Operators with Application to Parallel Optimization. PhD thesis, MIT (1989)
23. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for max. mon. operators. M. Prog. **55** (1992) 293–318
24. Michelot, C.: A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. J. Optim. Theory and Appl. **50**(1) (1986) 195–200
25. Dobson, D.C., Curtis, Vogel, R.: Iterative methods for total variation denoising. J. Sci. Comput **17** (1996) 227–238
26. Chambolle, A.: An algorithm for total variation minimization and applications. JMIV **20** (2004) 89–97
27. Chambolle, A.: Total variation minimization and a class of binary MRF models. In: EMMCVPR. Volume 3757. (2005) 136–152
28. Aujol, J.F.: Some algorithms for total variation based image restoration. CMLA Preprint (2008-05) (2008)
29. Chan, T.F., Golub, G.H., Mulet, P.: A nonlinear primal-dual method for total variation-based image restoration. J. Sci. Comput **20** (1999) 1964–1977
30. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. SIAM J. Multisc. Model. Sim. **4**(4) (2005) 1168–1200
31. Bresson, X., Chan, T.: Fast minimization of the vectorial total variation norm and applications to color image processing. Tech. Rep. 07-25, UCLA (2007)
32. Geman, D., Yang, C.: Nonlinear image recovery with halfquadratic regularization. IEEE Trans. Image Proc. **4**(7) (1995) 932–946
33. Cohen, L.: Auxiliary variables and two-step iterative algorithms in computer vision problems. JMIV **6**(1) (1996) 59–83
34. Strang, G.: The discrete cosine transform. SIAM Review **41**(1) (1999) 135–147
35. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for Markov random fields. In: ECCV. Volume 2. (2006) 19–26
36. Hintermüller, M., Stadler, G.: An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. J. Sci. Comput. **28**(1) (2006) 1–23