Non-local Total Generalized Variation for Optical Flow Estimation*

René Ranftl[1], Kristian Bredies[2], and Thomas Pock[1,3]

[1] Institute for Computer Graphics and Vision,
Graz University of Technology, Austria
[2] Institute for Mathematics and Scientific Computing,
University of Graz, Austria
[3] Safety & Security Department,
AIT Austrian Institute of Technology, Austria

Abstract. In this paper we introduce a novel higher-order regularization term. The proposed regularizer is a non-local extension of the popular second-order Total Generalized variation, which favors piecewise affine solutions and allows to incorporate soft-segmentation cues into the regularization term. These properties make this regularizer especially appealing for optical flow estimation, where it offers accurately localized motion boundaries and allows to resolve ambiguities in the matching term. We additionally propose a novel matching term which is robust to illumination and scale changes, two major sources of errors in optical flow estimation algorithms. We extensively evaluate the proposed regularizer and data term on two challenging benchmarks, where we are able to obtain state of the art results. Our method is currently ranked first among classical two-frame optical flow methods on the KITTI optical flow benchmark.

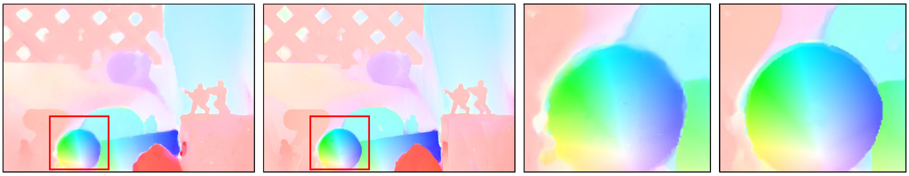
1 Introduction

Higher-order regularization has become increasingly popular for tackling correspondence problems like stereo or optical flow in recent years. This is not surprising since correspondences in real-world imagery can be modeled very well with the assumption of piecewise planar structures in the case of stereo estimation and piecewise affine motion in the case of optical flow.

Total Generalized Variation (TGV) [4], especially its second-order variant, has shown promising results as a robust regularization term. Consider for example the challenging KITTI Benchmark [9], where TGV-based optical flow models are currently among the top performing optical flow methods [3,23]. The merits of this regularization term are given by the fact that it is robust and allows for piecewise affine solutions. Moreover the regularization term is convex and

* René Ranftl and Thomas Pock acknowledge support from the Austrian Science Fund (FWF) under the projects No. I1148 and Y729. Kristian Bredies acknowledges support by the Austrian Science Fund special research grant SFB F32 "Mathematical Optimization and Applications in Biomedical Sciences".

D. Fleet et al. (Eds.): ECCV 2014, Part I, LNCS 8689, pp. 439–454, 2014.



(a) TGV (b) NLTGV (c) TGV (d) NLTGV

Fig. 1. Sample optical flow result from the Middlebury benchmark [1] using the proposed NLTGV regularizer compared to TGV. The regularizer is able to provide sharp and accurate motion boundaries and piecewise affine solutions.

a direct extension of the classical Total Variation semi-norm, which allows for easy integration into existing warping-based models. Note, however, that TGV suffers from the major drawback that it is local in its nature, *i.e.* only directly neighboring pixels influence the value of the regularization term, which may result in bad performance in areas where the data term is ambiguous. Moreover, purely TGV-based models are not able to accurately locate motion and depth discontinuities.

We propose a non-trivial non-local extension to the TGV regularization term, which is designed to remedy these problems. By incorporating larger neighborhoods into the regularizer and providing additional soft-segmentation cues, we are able to show increased performance in optical flow models. Our non-local regularizer remains convex and reduces to an anisotropic variant of the classical TGV regularizer for appropriately chosen neighborhoods, thus it is easy to integrate into existing frameworks. Figure 1 compares the proposed non-local Total Generalized Variation (NLTGV) to classical TGV. It can be seen that in both cases piecewise affine optical flow fields are obtained, but NLTGV results in significantly better localized motion boundaries.

A second important development, which is mainly driven by the recent availability of benchmarks featuring realistic data, is a strong interest in robust data terms. It is evident that in realistic scenarios, good optical flow estimates can only be obtained by a combination of a good regularization term as well as robust data terms. Rashwan *et al.* [20] incorporate dense HOG descriptors directly into the classical energy minimization framework in order to gain robustness against illumination changes, whereas [8] propose a simpler patch-based correlation measure, which is invariant to illumination and morphological changes. We again refer to the KITTI Benchmark, where many of the top-performing methods rely on variants of the Census transform for matching correspondences. The Census transform has shown to be robust to illumination changes both theoretically and in practice [11], which is especially important in realistic scenarios. Note, however, that an often overlooked additional source of errors are scale changes between images, which occur when motion along the optical axis is present in the scene. Classical patch-based data terms, such as the Census transform, fail in such scenarios, since the local appearance strongly changes in this case. To this end we introduce a novel dataterm, which is motivated by the Census transform, in order to gain robustness to scale changes, while still providing robustness to challenging illumination conditions. Our experiments show that using the proposed data term, we are able to obtain increased robustness in image sequences which feature scaling motions.

Related Work. Starting from the seminal work by Horn & Schunk [13], innumerable optical flow models have been proposed. An important development was the introduction of robust regularizers, specifically in the form of Total Variation regularization, and robust data terms [29]. Much research has been devoted to different aspects of this model, like edge-preserving regularization terms [26], or the robustness to large-displacement motions [28].

A non-local variant of Total Variation has been first introduced by Gilboa and Osher [10] for image and texture restoration problems. Werlberger *et al.* successfully showed that a smoothed variant of this regularizer can be used to incorporate soft-segmentation cues into motion estimation algorithms [25]. Sun *et al.* [21] arrived at a similar non-local model by formalizing a median filtering heuristic that is present in many successful optical flow models. Both models are computationally demanding if they are defined for large support window sizes, thus they are often constrained to small support windows. Krähenbühl *et al.* [15] showed how to approximately optimize optical flow models that incorporate non-local Total Variation in the presence of large support windows.

Models which incorporate TGV regularization have seen increasing success recently. Ranftl *et al.* [19] introduced a edge-aware TGV-based model with a Census data term for the task of stereo estimation. Similar to the popular LDOF [5] framework, Braux *et al.* [3] incorporate sparse feature matches into a TGV-based model in order to handle large displacements. Vogel *et al.* [23] also use a TGV-based model and investigate the influence of different robust data terms. These models currently define the state of the art on the KITTI optical flow benchmark.

All of these models use variants of the Census transform as data term in order to be robust against illumination changes, but surprisingly none of them explicitly consider scale changes. In the context of dense descriptor matching it was shown that it is possible to derive a "scaleless" version of the popular SIFT descriptor [12], which were integrated into the discrete SIFT-Flow framework [17]. Xu *et al.* incorporate scale estimation as an additional latent variable into a classical continuous optical flow model [27]. Since they model scale selection as a labeling problem, this model is computationally demanding. Finally, Kim *et al.* propose a locally adaptive fusion of different data costs [14], which in theory could also be used to remedy the negative influence of scale changes.

2 Preliminaries

We denote the optical flow field as $v = (v^1, v^2)^T : \Omega \to \mathbb{R}^2$ and the input images as $I_1, I_2 : \Omega \to \mathbb{R}$. A generic form of an optical flow energy takes the form

$$\min_v J(v^1) + J(v^2) + \lambda \int_\Omega \rho(x, v(x), I_1, I_2)\mathrm{d}x, \tag{1}$$

where $J(.)$ are the regularizers of the individual flow components, $\rho(x, v(x), I_1, I_2)$ is a matching term that gives the cost for warping I_1 to I_2 using the flow v and λ is a scalar regularization parameter.

In order to cope with the non-convexity of the matching term which arises from the warping operation and potentially from the function ρ, we follow the strategy of approximating the data term $\rho(x, v(x), I_1, I_2)$ using a second-order Taylor expansion [25] around some initial flow $v_0(x)$:

$$\begin{aligned}\rho(x, v(x)) \approx{}& \rho(x, v_0(x)) + (v(x) - v_0(x))^T \nabla \rho(x, v_0(x)) \\ &+ \tfrac{1}{2}(v(x) - v_0(x))^T (\nabla^2 \rho(x, v_0(x)))(v(x) - v_0(x)) = \hat{\rho}(x, v(x)), \end{aligned} \quad (2)$$

where we dropped the explicit dependence on I_1 and I_2 for notational simplicity. In contrast to the approach of linearizing the matching image [29], which leads to the classical optical flow constraint, this strategy allows to incorporate complex data terms into the model. As suggested in [25] we use a diagonal positive semi-definite approximation of the Hessian matrix $\nabla^2 \rho(x, v_0(x))$ in order to keep the approximation convex. The specific form of the regularization term and the matching term will be the subject of the next sections.

3 Non-local Total Generalized Variation

For clarity we focus on second-order regularization, since such regularizers have empirically shown to provide a good tradeoff between computational complexity and accuracy in correspondence problems.

Let $\Omega \subset \mathbb{R}^2$ denote the image domain and $u : \Omega \to \mathbb{R}$ be a function defined on this domain (*e.g.* one component of a flow field). The second-order Total Generalized Variation [4] of the function u is given by

$$\mathrm{TGV}^2(u) = \min_w \alpha_1 \int_\Omega |Du - w| + \alpha_0 \int_\Omega |Dw|, \quad (3)$$

where $w : \Omega \to \mathbb{R}^2$ is an auxiliary vector field, $\alpha_0, \alpha_1 \in \mathbb{R}^+$ are weighting parameters and the operator D denotes the distributional derivative, which is well-defined for discontinuous functions. An important property of this regularizer is that $\mathrm{TGV}^2(u) = 0$ if and only if u is a polynomial of order less than two [4], *i.e.* if u is affine. This explains the tendency of models, which incorporate this regularization term, to produce piecewise affine solutions. Note that the parameter α_1 is related to the penalization of jumps in u, whereas the parameter α_0 is related to the penalization of kinks, *i.e.* second-order discontinuities.

Non-local Total Variation [10] on the other hand can be defined as:

$$\mathrm{NLTV}(u) = \int_\Omega \int_\Omega \alpha(x, y)|u(x) - u(y)| dy dx. \quad (4)$$

Here, the support weights $\alpha(x, y)$ allow to incorporate additional prior information into the regularization term, *i.e.* $\alpha(x, y)$ can be used to strengthen the regularization in large areas, which is especially useful in the presence of ambiguous data terms. Variants of this regularizer have been successfully applied to the task of optical flow estimation [25,15,21].

Motivated by non-local Total Variation (4), Definition 1 introduces a non-local extension of the TGV^2 regularizer:

Definition 1. *Let $u : \Omega \to \mathbb{R}$, $w : \Omega \to \mathbb{R}^2$ and $\alpha_0, \alpha_1 : \Omega \times \Omega \to \mathbb{R}^+$ be support weights. We define the non-local second-order Total Generalized Variation regularizer $J(u)$ as*

$$J(u) = \min_w \int_\Omega \int_\Omega \alpha_1(x, y) |u(x) - u(y) - \langle w(x), x - y \rangle| \mathrm{d}y \mathrm{d}x + \sum_{i=1}^{2} \int_\Omega \int_\Omega \alpha_0(x, y) |w^i(x) - w^i(y)| \mathrm{d}y \mathrm{d}x, \tag{5}$$

where vector components are denoted by super-scripts, i.e. $w(x) = (w^1(x), w^2(x))^T$.

The reasoning behind this definition is as follows: Considering a point $x \in \Omega$, the expression $u(x) - \langle w(x), x - y \rangle$ defines a plane through the point $(x, u(x))$, with normal vector $(w(x), -1)^T$. Consequently the inner integral of the first expression,

$$\int_\Omega \alpha_1(x, y) |u(x) - u(y) - \langle w(x), x - y \rangle| \mathrm{d}y, \tag{6}$$

measures the total deviation of u from the plane at the point x, weighted by the support function α_1. The outer integral evaluates this deviation at every point in the image. This term can be understood as a linearization of u around a point x. Note that the linearization is not constant, *i.e.* as we are interested in a field w which minimizes the total deviations from the (in the continuous setting infinitely many) local planes, the normal vector $w(x)$ can vary, although not arbitrarily as the term

$$\sum_{i=1}^{2} \int_\Omega \int_\Omega \alpha_0(x, y) |w^i(x) - w^i(y)| \mathrm{d}y \mathrm{d}x \tag{7}$$

forces the field w to have low (non-local) total variation itself. Intuitively (5) assigns low values to functions u which can be well approximated by affine functions.

We now derive primal-dual and dual representations of (5), which will later serve as the basis for the optimization of functionals that incorporate this regularizer.

Proposition 1. *The dual of* (5) *is given by*

$$J(u) = \sup_{\substack{|p(x,y)|\leq\alpha_1(x,y)\\ |q^i(x,y)|\leq\alpha_0(x,y)}} \int_\Omega \left(\int_\Omega \{p(x,y) - p(y,x)\} \,\mathrm{d}y \right) u(x)\mathrm{d}x$$

$$s.t. \qquad \int_\Omega q^i(x,y) - q^i(y,x)\mathrm{d}y = \int_\Omega p(x,y)(x^i - y^i)\mathrm{d}y \quad \forall i \in \{1,2\} \tag{8}$$

Proof. Dualizing the absolute values in (5) yields

$$\begin{aligned} J(u) &= \min_w \sup_{|p(x,y)|\leq\alpha_1(x,y)} \int_\Omega \int_\Omega (u(x) - u(y) - \langle w(x), x-y\rangle) \cdot p(x,y)\mathrm{d}x\mathrm{d}y \\ &+ \sum_{i=1}^2 \sup_{|q^i(x,y)|\leq\alpha_0(x,y)} \int_\Omega \int_\Omega (w^i(x) - w^i(y)) \cdot q^i(x,y)\mathrm{d}x\mathrm{d}y \\ &= \min_w \sup_{\substack{|p(x,y)|\leq\alpha_1(x,y)\\ |q^i(x,y)|\leq\alpha_0(x,y)}} \int_\Omega \left(\int_\Omega \{p(x,y) - p(y,x)\} \,\mathrm{d}y \right) u(x)\mathrm{d}x \\ &+ \sum_{i=1}^2 \int_\Omega \left(\int_\Omega \left\{ q^i(x,y) - q^i(y,x) + p(x,y)(y^i - x^i) \right\} \mathrm{d}y \right) w^i(x)\mathrm{d}x. \end{aligned} \tag{9}$$

By taking the minimum with respect to w we arrive at the dual form. □

We will now show two basic properties of non-local Total Generalized Variation:

Proposition 2. *The following statements hold:*

1. *$J(u)$ is a semi-norm.*
2. *$J(u) = 0$ if and only if u is affine.*

Proof. To show the first statement, consider that the supremum in (8) is taken over linear functions with additional linear constraints on p and q. It is well-known that the supremum over linear functions is convex [2] . Since the constraints on p and q form a linear and thus convex set, $J(u)$ is convex. Moreover it is easy to see from (8) that $J(u)$ is positive one-homogeneous. As a consequence the triangle inequality holds, which establishes the semi-norm property.

In order to show the second statement, assume that u is affine, *i.e.* $u(x) = \langle a, x\rangle + b, \quad a \in \mathbb{R}^2$. By plugging into (5) it is easy to see that the minimum is attained at $w(x) = a$. As a consequence we have $J(u) = 0$. Conversely assume that $J(u) = 0$. In any case this requires that

$$\sum_{i=1}^2 \int_\Omega \int_\Omega \alpha_0(x,y)|w^i(x) - w^i(y)|\mathrm{d}y\mathrm{d}x = 0, \tag{10}$$

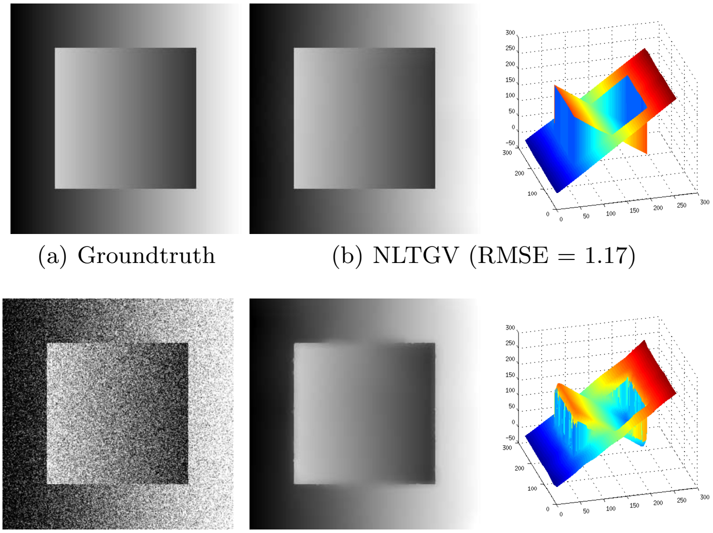
which implies that $w(x) = c \in \mathbb{R}^2$, $\forall x \in \Omega$. Consequently

$$\min_{c} \int_{\Omega} \int_{\Omega} \alpha_1(x, y) |u(x) - u(y) - \langle c, x - y \rangle | \mathrm{d}y \mathrm{d}x = 0, \tag{11}$$

if and only if $u(x)$ is of the form $u(x) = \langle a, x \rangle + b$ and hence affine. □

Since the properties in Proposition 2 are shared by TGV and the non-local TGV regularizer (NLTGV), it can be expected that both behave qualitatively similar when used in an energy minimization framework. The main advantage of NLTGV is the larger support size and the possibility to enforce additional prior knowledge using the support weights α_1 and α_0. This is especially advantageous for optical flow estimation, where support weights can be readily computed from a reference image, in order to allow better localization of motion boundaries and resolve ambiguities. Akin to [25] the support weights α_1 and α_0 can be used to incorporate soft-segmentation cues into the regularizer, *e.g.* in the case of optical flow estimation it is possible to locally define regions which are forced to have similar motion based on the reference image.

Figure 2 shows a synthetic experiment which demonstrates the qualitative behavior of NLTGV. We denoise a piecewise linear function using a quadratic data term with TGV and NLTGV, respectively. We assume prior knowledge of jumps in order to compute the support weights and set $\alpha_1(x, y) = 1$ if there is no discontinuity between x and y and $\alpha_1(x, y) = 0.1$ otherwise. Support weights outside of a 5×5 window were set to zero. While prior knowledge of jumps is not



(a) Groundtruth

(b) NLTGV (RMSE = 1.17)

(c) Noisy

(d) TGV (RMSE = 5.59)

Fig. 2. Comparison of NLTGV and TGV for denoising a synthetic image. NLTGV is able to perfectly reconstruct the groundtruth image. TGV tends to oversmooth jumps.

available in real denoising problems, similar support weights can be easily derived in optical flow estimation from the input images. It can be seen that NLTGV nearly perfectly reconstructs the original image, while TGV has problems with accurate localization of the discontinuities.

4 Scale-Robust Census Matching

The Census transform is a popular approach to gain robustness against illumination changes in optical flow. The principal idea is to generate a binary or ternary representation, called Census signature, of an image patch and measures patch similarity using the Hamming distance between Census signatures.

Let us define the per-pixel Census assignment function for an image $I : \Omega \to \mathbb{R}$:

$$C_\varepsilon(I, x, y) = \operatorname{sgn}(I(x) - I(y))\mathbb{1}_{|I(x)-I(y)|>\varepsilon}, \tag{12}$$

which assigns to the pixel at location y one of the values $\{-1, 0, 1\}$ based on the value of the pixel x. Given two images I_1, I_2 and a flow field $v : \Omega \to \mathbb{R}^2$, the Census matching cost of the flow v is defined via the Hamming distance of the two strings as

$$\rho_c(x, v(x), I_1, I_2) = \int_\Omega \mathbb{1}_{C_\varepsilon(I_1,x,y)\neq C_\varepsilon(I_2,x+v(x),y+v(x))}\mathcal{B}(x-y)\mathrm{d}y, \tag{13}$$

where $\mathcal{B}$ denotes a box filter, which defines the size of the matching window.

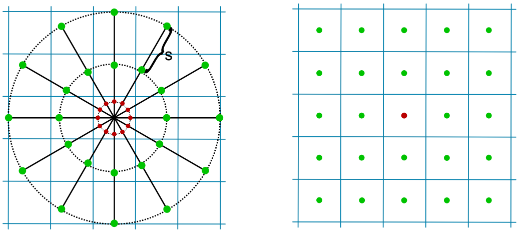
Note that classical patch-based matching approaches are problematic when scale changes between two images occur, since the patch in the first image will capture different features than the patch in the second image. If one knew the amount of scale change, a simple remedy to this problem would be to appropriately rescale the patch, such that the local appearance is again the same. Unfortunately the scale change in optical flow estimation is unknown a-priori.

To this end we draw ideas from SIFT descriptor matching under scale changes in order to alleviate these problems: Consider SIFT descriptors h^1 and h^2 computed from two images I_1 and I_2 at points p^1 and p^2 respectively. Hassner *et al.* [12] showed that if descriptors are sampled at different scales s_i and the "min-dist" measure, which is defined as

$$\min_{i,j} dist(h^1_{s_i}, h^2_{s_j}), \tag{14}$$

is used as matching score, it is possible to obtain accurate matches even under scale changes. Since SIFT descriptors are based on distributions of image gradients and [11] has shown a strong relationship of the Census transform to an anisotropic gradient constancy assumptions, it is reasonable to assume that a similar strategy might be applicable to Census transform matching.

We define a variant of the Census transform, which is easily amenable for multi-scale resampling, by using radial sampling instead of a window-based sampling strategy. An example of this sampling strategy is shown in Figure 3. We



(a) SCensus (b) Census

Fig. 3. Example of the proposed sampling strategy analogous to a 5x5 census transform. The center value is computed by averaging the sampling positions on the inner most ring (red). A ternary string of length 24 is generated from the sampling positions on the outer rings (green). (Best viewed in color).

sample radially around the center point. Samples from the inner ring are averaged and serve as the basis value for generating the Census string, *i.e.* the average takes the role of the center pixel when compared to the standard Census transform. In order to generate the Census string, the gray values of samples on the outer ring are compared to the average value. All samples are extracted using bilinear interpolation, whenever a sampling point is not in the center of a pixel. This strategy allows simple rescaling of the descriptor, which is important for an efficient implementation. Note that this radial sampling shares similarities to Local Binary Patterns [18]. Formally, we fix some radial discretization step $\theta = \frac{2\pi}{K}$ and a radius r and introduce scale depended coordinates $\hat{x} = (\hat{x}^1, \hat{x}^2)^T$

$$\hat{x}^1(k,s,r) = x^1 + rs\cos(k\theta), \qquad \hat{x}^2(k,s,r) = x^2 + rs\sin(k\theta) \tag{15}$$

We define the difference between the average value of the inner ring $r_i = \frac{s}{4}$ and the l-th sample from an outer ring r as

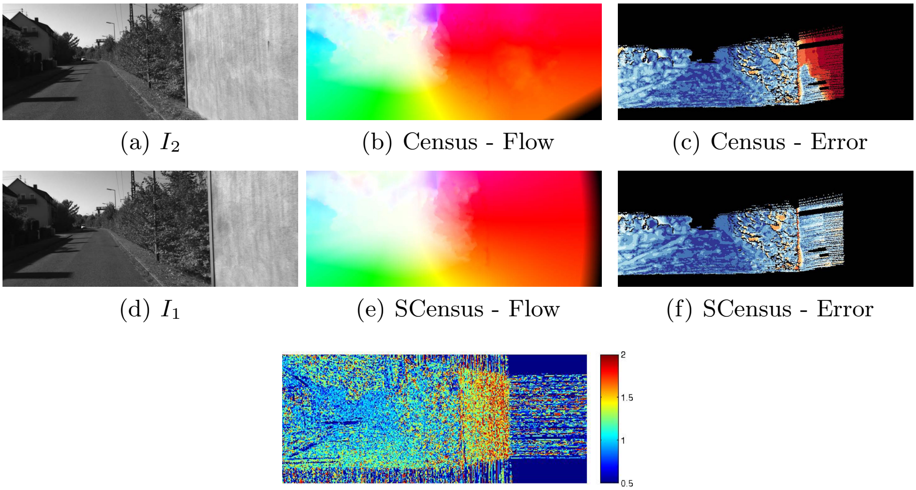
$$f(I,x,l,s,r) = \frac{1}{K}\sum_{k=1}^{K}(G_s * I)(\hat{x}(k,s,\tfrac{s}{4})) - (G_s * I)(\hat{x}(l,s,r)), \tag{16}$$

where G_s denotes a Gaussian kernel with variance s. Analogous to the Census assignment function (12) we define the scale-dependent Census assignment function as

$$C_\varepsilon^s(I,x,l,r) = \operatorname{sgn}(f(I,x,l,s,r))\mathbb{1}_{|f(I,x,l,s,r)|>\varepsilon}, \tag{17}$$

This definition allows to compare descriptors at different scales s_1 and s_2 using the Hamming distance:

$$\rho_{s_2}^{s_1}(x,v(x),I_1,I_2) = \sum_{l=1}^{L}\sum_{r=1}^{R}\mathbb{1}_{C_\varepsilon^{s_1}(I_1,x,l,r)\neq C_\varepsilon^{s_2}(I_2,x+v(x),l,r)}. \tag{18}$$


(a) I_2 (b) Census - Flow (c) Census - Error

(d) I_1 (e) SCensus - Flow (f) SCensus - Error

(g) Selected Scale

Fig. 4. Example behaviour of the Census dataterm and the scale-robust Census dataterm. The wall to the right undergoes a strong scale change. (b)-(c): Census fails in these areas. (e)-(f): Using scale-robust Census we are able to find a correct flow field. (g) shows the scale that was locally selected by the data term. (Best viewed in color).

By introducing the "min-dist" measure we finally arrive at the scale-robust Census data term:

$$\rho(x, v(x), I_1, I_2) = \min_{s_1, s_2} \rho_{s_2}^{s_1}(x, v(x), I_1, I_2). \tag{19}$$

While this data term is highly non-linear and non-convex, it can still be easily integrated into our continuous model using the convex quadratic approximation (2).

In practice we fix the scale in the first first image to the original scale and compute $\rho_{s_2}^1$ for a number of scales s_2. Note that this definition is slightly biased toward forward motion, but is also able to handle moderate scale changes in the other direction.

Figure 4 shows the qualitative behavior of the proposed data term in areas that undergo a strong scale change. It can be seen that the proposed data term is able to successfully choose the correct scale on many points, which allows the global model to achieve accurate results.

5 Discretization and Minimization

For minimization we use the preconditioned primal-dual scheme [7]. We discretize (1) on the regular rectangular pixel grid of size $M \times N$ and use the index $1 \leq i \leq MN$ to refer to individual pixels in this grid. Let $v^i \in \mathbb{R}^2$ denote the flow at the i-th pixel, which is at the location $l_i = (x^1(i), x^2(i))^T$. In order to allow for a simpler notation, we introduce a signed distance matrix

$$D_{ij} = \begin{pmatrix} d^1_{ij} & d^2_{ij} & 0 & 0 \\ 0 & 0 & d^1_{ij} & d^2_{ij} \end{pmatrix} \in \mathbb{R}^{2\times 4},$$

with $d_{ij} = (d^1_{ij}, d^2_{ij})^T = l_j - l_i$. Let $p^{ij} \in \mathbb{R}^2$ and $q^{ij} \in \mathbb{R}^4$ be the dual variable associated to the connection of pixels i and j. The discretized model can be written in its primal-dual formulation as

$$\min_{v,w} \max_{\substack{\|p^{ij}\|_\infty \le \alpha_1^{ij} \\ \|q^{ij}\|_\infty \le \alpha_0^{ij}}} \sum_i \sum_{j>i} \left[(v^i - v^j + D_{ij} w^i) \cdot p^{ij} + (w^i - w^j) \cdot q^{ij} \right] + \lambda \sum_i \hat\rho(i, v^i). \tag{20}$$

Remark 1. In order to prevent double counting of edges we set the support weights in (20) to zero for all $y^1(i) \le x^1(i)$ or $(y^2(i) \le x^2(i)) \wedge (y^1(i) \le x^1(i))$.

Using (9) we can derive the optimization scheme:

$$\begin{cases} p^{ij}_{n+1} & = \max(-\alpha_1^{ij}, \min(\alpha_1^{ij}, p^{ij}_n + \sigma_p(\bar v^i_n - \bar v^j_n + D_{ij} \bar w^i_n) \\ q^{ij}_{n+1} & = \max(-\alpha_0^{ij}, \min(\alpha_0^{ij}, q^{ij}_n + \sigma_q(\bar w^i_n - \bar w^j_n))) \\ v^i_{n+1} & = \mathrm{prox}_{\tau_v \lambda \hat\rho}(v^i_n - \tau_v \sum_{j>i}(p^{ij}_{n+1} - p^{ji}_{n+1})) \\ w^i_{n+1} & = w^i_n - \tau_w \sum_{j>i}(q^{ij}_{n+1} - q^{ji}_{n+1} + D^T_{ij} p^{ij}_{n+1}) \\ \bar v^i_{n+1} & = 2v^i_{n+1} - v^i_n \\ \bar w^i_{n+1} & = 2w^i_{n+1} - w^i_n \end{cases}$$

where minima and maxima are taken componentwise. The proximal operator $\mathrm{prox}_{t\hat\rho}(\hat u)$ with respect to the quadratic approximation of the data term is given by

$$\mathrm{prox}_{t\hat\rho}(\hat v^i) = (\nabla^2 \rho(v^i_0) + \tfrac{1}{t} I)^{-1} (\tfrac{1}{t} \hat v^i - \nabla \rho(v^i_0) + \nabla^2 \rho(v^i_0) v_0). \tag{21}$$

We compute support weights based on color similarities and spatial proximity:

$$\alpha_1^{ij} = \frac{1}{Z^i} \exp(-\tfrac{\|I^i_1 - I^j_1\|}{w_c}) \exp(-\tfrac{\|l_j - l_i\|}{w_p}), \quad \alpha_0^{ij} = c\alpha_1^{ij}, \tag{22}$$

where w_c and w_p are user-chosen parameters that allow to weight the influence of the individual terms and Z^i ensures that the support weights sum to one. Note that in practice we constrain the influence of the non-locality in a window of size $2w_p + 1$ in order to keep optimization tractable (e.g. weights outside the window are set to zero, which allows to drop corresponding dual variables from the optimization problem). Figure 5 shows the influence of the parameters w_p and and w_c on the average endpoint error (EPE), evaluated on the Middlebury training set [1]. It can be seen that larger spatial influence results in lower EPE, whereas a too large color similarity parameter results in oversmoothing and consequently yields higher EPE.

As is common, the optimization is embedded into a coarse-to-fine warping framework in order to cope with large motions.

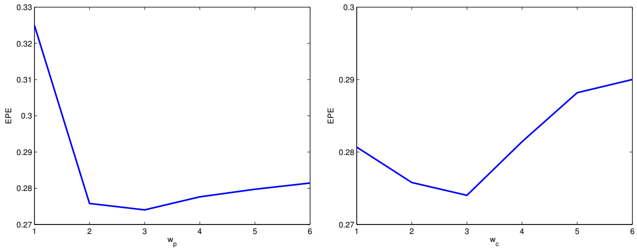


Fig. 5. Influence of the spatial proximity parameter w_p an the color proximity parameter w_c on EPE evaluated on the Middlebury training set

6 Experiments

In this section we evaluate the performance of the proposed model on two challenging data sets. The model was implemented using CUDA; all experiments were conducted on a Geforce 780Ti GPU. We use a scale factor of 0.8 for the coarse-to-fine pyramid and 15 warps per pyramid level. For the scale-robust data term we evenly sample 7 scales between 0.5 and 2 in both image. We fix $w_p = 2$, which gives a good trade-off between accuracy and computational complexity. The remaining parameters were adapted for each benchmark individually.

KITTI Benchmark. The KITTI Benchmark [9] is composed of real-world images taken from an automotive platform. The data set is split into a training set and a test set of 194 images each. We use the training set, where groundtruth optical flow is available, to show the influence of non-local TGV as well as the scale-robust data term. As a baseline model we use standard TGV with the Census term (TGV-C), as it has been shown that this combination already works well on this dataset. We compare different combinations of regularizers and data terms: Standard TGV, non-local TV, as defined in (4), and NLTGV. The suffixes -C and -SC denote Census and scale-robust Census, respectively.

We use a small subset of the training set (20% of the images) to find optimal parameters for each method using grid-search. The Census and NLTGV window sizes were set to 5×5. Since the groundtruth flow fields in this data set are not

Table 1. Average error in % for different models and different error thresholds on the KITTI NOC-training set

	TGV-C	NLTV-C	NLTGV-C	TGV-SC	NLTV-SC	NLTGV-SC
$2px$	12.86	12.38	7.58	11.73	11.29	**7.35**
$3px$	10.38	9.59	5.74	9.19	8.57	**5.50**
$4px$	8.99	8.27	4.90	7.87	7.30	**4.59**
$5px$	8.03	7.48	4.34	6.97	6.53	**4.00**

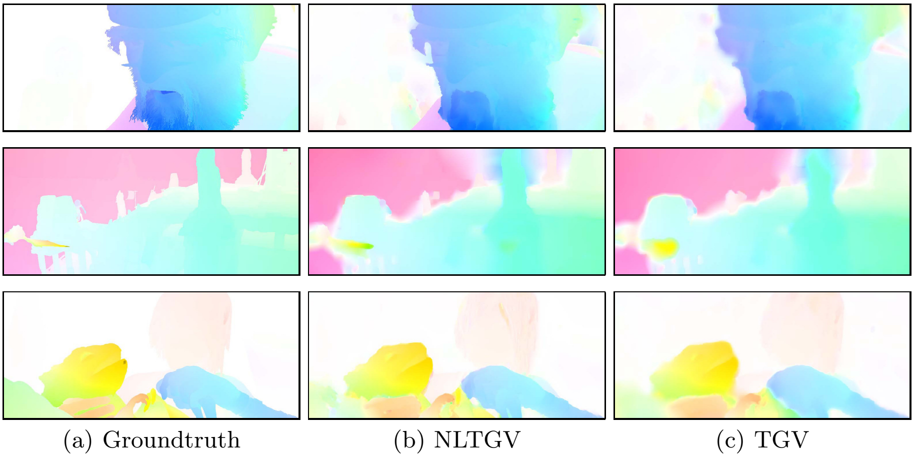


Fig. 6. Comparison between NLTGV and TGV on the Sintel Benchmark

pixel-accurate, we follow the officially suggested methodology of evaluating the percentage of pixels, which have endpoint error above some threshold [9].

Table 1 shows a comparison of TGV and NLTV to NLTGV, as well as the influence of the scale-robust Census data term. TGV and NLTV perform similar, which is in accordance to the results of similar NLTV-based models on this dataset (*cf.* [22]). NLTGV gives a significantly lower error with both data terms. This can be attributed to more accurate motion boundaries and a better behaviour in occluded and ambiguous areas. Using the scale-robust Census data term additionally lowers the error for both models, with NLTGV-SC giving the lowest overall error. Table 2 shows results on the test set of this benchmark, where our method is currently ranked first among two-frame optical flow methods.

Table 2. Average error on the KITTI test set for error thresholds $3px$ and $2px$. Suffixes "Noc" and "All" refer to errors evaluated in non-occluded and all regions, respectively. Methods "DDR-DF" and "EpicFlow" were unpublished at the time of writing. We show the six best-performing two-frame optical flow methods.

	Out-Noc [%]		Out-All [%]		Avg-Noc [px]	Avg-All [px]	Runtime [s]
	3px	2px	3px	2px			
NLTGV-SC	**5.93**	**7.64**	**11.96**	**14.55**	1.6	3.8	16
DDR-DF	6.03	8.23	13.08	16.01	1.6	**2.7**	60
TGV2ADCS [3]	6.20	8.04	15.15	17.87	1.5	4.5	12
DataFlow [23]	7.11	9.16	14.57	17.41	1.9	5.5	180
EpicFlow	7.19	9.53	16.15	19.47	**1.4**	3.7	15
DeepFlow [24]	7.22	9.31	17.79	20.44	1.5	5.8	17

Table 3. Average EPE for a selection of different models on the Sintel test set. The columns “sA-B” refer to EPE over regions with velocities between A and B.

Rank	Method	EPE all	s0-10	s10-40	s40+
1	EpicFlow	6.469	1.180	4.000	38.687
4	DeepFlow [24]	7.212	1.284	4.107	44.118
21	NLTGV-SC	8.746	1.587	4.780	53.860
23	DataFlow [23]	8.868	1.794	5.294	52.636
28	NLTV-SC	9.855	1.202	4.757	64.834

Sintel Benchmark. The synthetic Sintel Benchmark [6] features large motion, challenging illumination conditions and specular reflections. In our evaluation we use the “final” sequence, which additionally contains motion blur and atmospheric effects. We use two image pairs from each subsequence of the training set to set the parameters and report the average endpoint error as error measure.

Table 3 show results on the Sintel test set. We see an improvement over the TGV-based model [23] and an NLTV-based model (NLTV-SC). The most critical regions for the overall error are high-velocity regions, which are problematic in purely coarse-to-fine-based methods. Hence, it is not surprising that methods which integrate some form of sparse prior matching [24,16] fair better than classical coarse-to-fine-based approaches on this dataset. Note that a-priori matches could be easily integrated into our model [3]. We leave such an extension for future work. Finally, Figure 6 shows a qualitative comparison between TGV and NLTGV on this benchmark.

7 Conclusion

In this paper we have introduced a novel higher-order regularization term for variational models, called non-local Total Generalized Variation. The principal idea of this regularizer is to measure deviations of a function from local linear approximations, where an additional spatial smoothness assumption is imposed onto the linear approximations. The proposed regularization term allows for piecewise affine solutions and is able to incorporate soft-segmentation cues, which is especially appealing for tasks like optical flow estimation and stereo. Additionally, we introduced a novel data term for optical flow estimation, which is robust to scale and illumination changes, as they frequently occur in optical flow imagery. Our experiments show that an optical flow model composed of non-local Total Generalized Variation together with the proposed scale robust data term is able to significantly improve optical flow accuracy.

References

1. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision 92(1), 1–31 (2011)

2. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York (2004)
3. Braux-Zin, J., Dupont, R., Bartoli, A.: A general dense image matching framework combining direct and feature-based costs. In: International Conference on Computer Vision, ICCV (2013)
4. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM Journal on Imaging Sciences 3(3), 492–526 (2010)
5. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(3), 500–513 (2011)
6. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012)
7. Chambolle, A., Pock, T.: A first-order primal-dual algorithm or convex problems with applications to imaging. Journal of Mathematical Imaging and Vision 40(1), 120–145 (2011)
8. Demetz, O., Hafner, D., Weickert, J.: The complete rank transform: A tool for accurate and morphologically invariant matching of structures. In: British Machine Vision Conference, BMVC (2013)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition, CVPR (2012)
10. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. Multiscale Modeling & Simulation 7(3), 1005–1028 (2008)
11. Hafner, D., Demetz, O., Weickert, J.: Why is the census transform good for robust optic flow computation? In: Kuijper, A., Bredies, K., Pock, T., Bischof, H. (eds.) SSVM 2013. LNCS, vol. 7893, pp. 210–221. Springer, Heidelberg (2013)
12. Hassner, T., Mayzels, V., Zelnik-Manor, L.: On sifts and their scales. In: Conference on Computer Vision and Pattern Recognition, CVPR (2012)
13. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artifical Intelligence 17, 185–203 (1981)
14. Kim, T.H., Lee, H.S., Lee, K.M.: Optical flow via locally adaptive fusion of complementary data costs. In: International Conference on Computer Vision, ICCV (2013)
15. Krähenbühl, P., Koltun, V.: Efficient nonlocal regularization for optical flow. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 356–369. Springer, Heidelberg (2012)
16. Leordeanu, M., Zanfir, A., Sminchisescu, C.: Locally affine sparse-to-dense matching for motion and occlusion estimation. In: International Conference on Computer Vision, ICCV (2013)
17. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(5), 978–994 (2011)
18. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition 29(1), 51–59 (1996)
19. Ranftl, R., Gehrig, S., Pock, T., Bischof, H.: Pushing the Limits of Stereo Using Variational Stereo Estimation. In: Intelligent Vehicles Symposium (2012)

20. Rashwan, H.A., Mohamed, M.A., García, M.A., Mertsching, B., Puig, D.: Illumination robust optical flow model based on histogram of oriented gradients. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 354–363. Springer, Heidelberg (2013)
21. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2432–2439 (2010)
22. Sun, D., Roth, S., Black, M.: A quantitative analysis of current practices in optical flow estimation and the principles behind them. International Journal of Computer Vision 106(2), 115–137 (2014)
23. Vogel, C., Roth, S., Schindler, K.: An evaluation of data costs for optical flow. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 343–353. Springer, Heidelberg (2013)
24. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: DeepFlow: Large displacement optical flow with deep matching. In: International Conference on Computer Vision, ICCV (2013)
25. Werlberger, M., Pock, T., Bischof, H.: Motion estimation with non-local total variation regularization. In: Conference on Computer Vision and Pattern Recognition, CVPR (2010)
26. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic huber-l1 optical flow. In: British Machine Vision Conference, BMVC (2009)
27. Xu, L., Dai, Z., Jia, J.: Scale invariant optical flow. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 385–399. Springer, Heidelberg (2012)
28. Xu, L., Jia, J., Matsushita, Y.: Motion detail preserving optical flow estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(9), 1744–1757 (2012)
29. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713, pp. 214–223. Springer, Heidelberg (2007)