An Evaluation of Data Costs for Optical Flow

Christoph Vogel [1], Stefan Roth [2] and Konrad Schindler [1]

[1] Photogrammetry and Remote Sensing, ETH Zurich, Switzerland
[2] Department of Computer Science, TU Darmstadt, Germany

Abstract. Motion estimation in realistic outdoor settings is significantly challenged by cast shadows, reflections, glare, saturation, automatic gain control, *etc*. To allow robust optical flow estimation in these cases, it is important to choose appropriate data cost functions for matching. Recent years have seen a growing trend toward patch-based data costs, as they are already common in stereo. Systematic evaluations of different costs in the context of optical flow have been limited to certain cost types, and carried out on data without challenging appearance. In this paper, we contribute a systematic evaluation of various pixel- and patch-based data costs using a state-of-the-art algorithmic testbed and the realistic KITTI dataset as basis. Akin to previous findings in stereo, we find the Census transformation to be particularly suitable for challenging real-world scenes.

1 Introduction and Related Work

Optical flow estimation has come far since the pioneering works of Lucas/Kanade [16] and Horn/Schunck [14]. Modern optical flow algorithms are reaching a point where they become suitable for deployment in real-world vision systems, *e.g*. [17]. Still, most state-of-the-art methods continue to be variants of the continuous energy minimization framework of [14], *i.e*. they formulate an energy that aggregates data and smoothness costs over the image as a function of the flow field, and then seek to minimize it.

Aside from the enormous growth in compute power, various factors have driven the progress of optical flow methods: *(i)* Every optical flow approach needs to incorporate prior assumptions due to the ill-posedness of estimating 2D motion vectors from changes in image brightness (the so-called aperture problem). Robust [3, 6] and non-local regularizers [24, 27] have greatly improved the ability to estimate flow in areas where the observed image data is weak or ambiguous. *(ii)* Also the data costs (respectively likelihoods) have evolved to better deal with noise, lighting changes, *etc*. While optical flow methods were originally based on the brightness constancy assumption [14], and later robust versions of it [3], there has been a recent trend towards more expressive, usually patch-based cost functions [17, 23], as is already common in stereo matching [13]. *(iii)* Sophisticated optimization schemes [7, 8, 29] have made flow estimation much more efficient. Moreover, other algorithmic advances, such as pre-filtering or outlier removal [26], have further increased the robustness of the estimate.

Designing robust optical flow approaches that work well in realistic settings, such as challenging outdoor scenes with cast shadows, reflections, glare, saturation, automatic gain control, *etc*., requires carefully choosing all components of the approach. Not only for this reason, benchmarking has a long tradition in optical flow research [1, 2]. Yet, the

focus of quantitative evaluations typically lies on showing the relative performance of entire methods, which differ from one another in terms of the regularizer, the mathematical and algorithmic framework, and sometimes also the data cost. Systematic studies of the benefit of the individual components are scarce, and have been largely limited to choices related to the prior and algorithmic aspects [24]. Systematic evaluations of data costs, especially patch-based ones, are rare and limited to a few metrics [23]. We address this by evaluating different cost functions for optical flow estimation in a consistent framework, in order to isolate the contribution of the data cost.

We focus on global optical flow methods and use standard total variation [29] and more powerful total generalized variation [5] regularizers as testbed, along with primal-dual optimization [8]. Our evaluation relies on two suitable benchmark datasets, the now classic *Middlebury* benchmark [1], and the more recent *KITTI* dataset [9]. Our testbed matches the performance of other implementations on Middlebury, thus suggesting that it is representative of the current state-of-the-art. We make the following contributions: *(i)* We conduct an evaluation and comparison of several data terms – brightness constancy with and without prefiltering [26], normalized cross correlation [27], mutual information [18], and the census transform [17] – in a consistent framework; *(ii)* we introduce a variant of the census transform that allows it to be embedded in gradient-based inference schemes; *(iii)* in the course of the evaluation, we achieve some of the most accurate results reported to date on the challenging *KITTI* dataset.

2 Optical Flow Testbed

Given two subsequent images from a video sequence $\mathbf{I}_0, \mathbf{I}_1 : \Omega \to \mathbb{R}^+$, defined over an image domain Ω, we aim to estimate the optical flow $\mathbf{v} : \Omega \to \mathbb{R}^2$ composed of a horizontal and vertical component, $\mathbf{v} = (u, v)$. More precisely, we aim to compute the 2D motion field, such that the image points $I_0(\mathbf{p})$ and $I_1(\mathbf{p}+\mathbf{v}_\mathbf{p})$ are observations of the same physical scene point. To assess how well a given motion field explains the image data, a similarity measure $E_D(I_0, I_1, \mathbf{v})$ over matching pixels, termed "data cost", is defined. The simplest one is the pixelwise brightness constancy assumption (BCA).

Only demanding data fidelity leaves optical flow estimation ill-posed. This is resolved by imposing prior assumptions, such as the smoothness of the motion field, through a regularizer $E_S(\mathbf{v})$. Most modern optical flow algorithms minimize a global cost function consisting of a weighted combination of both energies:

$$\lambda E_D(I_0, I_1, \mathbf{v}) + E_S(\mathbf{v}) \quad \to \min_{\mathbf{v}} \tag{1}$$

Since the data cost is not convex due to the inherent ambiguity in matching, optimization typically proceeds in a coarse-to-fine manner to cope with large displacements [6].

A popular choice for $E_S(\mathbf{v})$ is the Total Variation [21, 29] regularizer defined as $\mathrm{TV}(\mathbf{v}) = \mathrm{TV}(u) + \mathrm{TV}(v) = \int_\Omega |\nabla u| + |\nabla v| \, \mathrm{d}\mathbf{p}$. Later, [5] generalized TV to Total Generalized Variation (TGV). While TV favors piecewise constant flow, TGV allows for solutions of a higher polynomial degree: a regularizer TGV^k_α of order k assigns zero energy to polynomials of order $k-1$. In its primal form, TGV^2_α can be written as

$$\mathrm{TGV}^2_\alpha(u) = \min_w \left(\alpha_1 \int_\Omega |\nabla u - w| \, \mathrm{d}\mathbf{p} + \alpha_0 \int_\Omega |\nabla w| \, \mathrm{d}\mathbf{p} \right). \tag{2}$$

As for TV, we define $\mathrm{TGV}^2_\alpha(\mathbf{v}) = \mathrm{TGV}^2_\alpha(u) + \mathrm{TGV}^2_\alpha(v)$. For Middlebury, which has largely fronto-parallel scenes, we found TV to work better, and consequently use it in our evaluation. On the contrary, the geometric layout of street scenes causes approximately piecewise linear flow fields when viewed by a forward-moving observer. For KITTI we thus use TGV^2_α, which we found to clearly outperform TV on that dataset.

Estimation algorithm. To minimize TV and TGV we follow [8], which proposes an efficient primal-dual solver for problems of the type

$$\min_{\mathbf{v}} F(D\mathbf{v}) + G(\mathbf{v}). \tag{3}$$

Here, it is assumed that the optimization (*i.e*. the flow field $\mathbf{v}$) is discretized to a regular $M \times N$ pixel grid. To see how this applies to optical flow, we consider the case of TV, to simplify notation. The derivation for TGV is similar.

We define $D : V \to Y$ to be a linear operator that for a (vectorized) flow field $\mathbf{v} \in V = \mathbb{R}^{2NM}$ yields the horizontal and vertical flow derivatives, approximated by finite differences, and concatenated into a vector in $Y = \mathbb{R}^{4NM}$. We can now estimate optical flow according to Eq. (1) by setting $G = \lambda E_D$ and $F(D\mathbf{v}) = \sum_{\mathbf{p}} ||(D\mathbf{v})_{\mathbf{p}}|| = \sum_{\mathbf{p}} \sqrt{\nabla u^t_{\mathbf{p}} \nabla u_{\mathbf{p}} + \nabla v^t_{\mathbf{p}} \nabla v_{\mathbf{p}}}$. If both F and G are convex mappings, the primal formulation can be recast as a saddle point problem, via the Legendre-Fenchel transform:

$$\min_{\mathbf{v}\in V} \max_{\mathbf{y}\in Y} \langle D\mathbf{v}, \mathbf{y}\rangle - F^*(\mathbf{y}) + G(\mathbf{v}), \tag{4}$$

where $F^*(\mathbf{y}) = \sup_{\mathbf{z}\in Y} \langle \mathbf{z}, \mathbf{y}\rangle - F(\mathbf{z})$ is the conjugate function of F. Starting from $\mathbf{v}_0 \in V, \mathbf{y}_0 \in Y$, Eq. (4) can now be minimized efficiently [8] by iterating over k and updating $(\mathbf{v}_k, \mathbf{y}_k)$ according to

$$\mathbf{v}^{k+1} = (I + \tau\partial G)^{-1}\big(\mathbf{v}^k - \tau D^T \mathbf{y}^k\big) \tag{5a}$$

$$\mathbf{y}^{k+1} = (I + \sigma\partial F^*)^{-1}\big(\mathbf{y}^k + \sigma D(2\mathbf{v}^{k+1} - \mathbf{v}^k)\big), \tag{5b}$$

with $\sigma^{-1}\cdot\tau^{-1} \leq ||D||^2 = 8$ to ensure convergence (in practice normally $\sigma = \tau = 1/\sqrt{8}$). The proximal operator $(I + \tau\partial G)^{-1}(\hat{\mathbf{v}})$, needed to solve Eq. (5a), is then given as

$$(I + \tau\partial G)^{-1}(\hat{\mathbf{v}}) := \arg\min_{\mathbf{v}} \frac{1}{2\tau}||\mathbf{v} - \hat{\mathbf{v}}||^2 + G(\mathbf{v}). \tag{6}$$

The proximal operator for F^* is defined similarly and given by a pixelwise projection

$$\big((I + \sigma\partial F^*)^{-1}(\mathbf{y})\big)_{\mathbf{p}} = \frac{\mathbf{y}_{\mathbf{p}}}{\max(1, ||\mathbf{y}_{\mathbf{p}}||)}, \tag{7}$$

onto the unit ball. The algorithm for TGV^2_α proceeds in the same way, using a different linear operator, see *e.g*. [20]. Note that in this case the update step width can be set individually per pixel ($\tau_{\mathbf{p}}$, $\sigma_{\mathbf{p}}$), following [19].

Datasets. The *Middlebury* dataset [1] has been the standard optical flow benchmark for several years. It is widely used (90 entries as of July 2013) and has very precise, dense ground truth. Its main limitations are the somewhat artificial scenes, with strong

contrasts, well-saturated colors, and little deviation from diffuse surface shading. Also, the scenes have a bias towards piecewise planarity and fronto-parallel depth layers.

The more recent *KITTI* dataset [9] is recorded outdoors from a moving vehicle. The images exhibit more realistic imaging conditions, with cast shadows, glare, specular reflections, changes in camera gain, *etc*., complicating flow estimation. Weaknesses include the ground truth from a laser scanner, which is only available at a sparse set of points. These points are irregularly distributed with a noticeable "near-field bias": Surfaces closer to the camera have many more ground truth points (thus influence on the error); also the maximum depth in the field of view regularly exceeds the scanner's depth range, and the scenes are almost completely static except for the ego-motion.[1]

3 Evaluated Data Costs

We first introduce the different data terms independent of a particular optimization framework, and defer details on how to embed them in the specific algorithm used here (*i.e*. how to efficiently solve the corresponding proximal maps). Note that for all patch-based data terms we warp I_1 based on the current motion field, and evaluate the similarity w.r.t. the warped image. This is in contrast to approaches that assume fronto-parallel patch motion, *cf*. also [4]. Even though interesting, we do not consider explicit models of brightness changes, *e.g*. [11], in this necessarily limited study.

BCA. Probably the simplest and most common data cost embodies the brightness constancy assumption (BCA), *i.e*. penalizing grayvalue changes of a moving surface point:

$$\mathrm{BCA}(\mathbf{v}) = \int_{\Omega} |I_0(\mathbf{p}) - I_1(\mathbf{p}+\mathbf{v})| \, \mathrm{d}\mathbf{p}. \tag{8}$$

It is common to linearize the BCA (*cf*. Sec. 3.1) and employ a robust penalizer, *e.g*. L_1 (Eq. 8) or a differentiable approximation. In real scenes the BCA is often violated due to non-Lambertian reflectance, varying illumination, *etc*. To mitigate the impact of brightness changes, one can apply structure-texture (TV-L_2) decomposition (STT, [21]) as a preprocessing step, *cf*. [26]. The images are separated into a piecewise constant "structure" part and a high-frequency "texture" part, from which the flow is estimated.

NCC. Another popular data cost is the normalized cross correlation (NCC, [23, 27]). For a single pixel location $\mathbf{p}$, the NCC is defined as the integral over a small neighborhood $\mathcal{N}(\mathbf{p})$, specified with a box filter B,

$$\mathrm{NCC}(\mathbf{p},\mathbf{v}) = \int_{\Omega} \frac{\big(I_0(\mathbf{y})-\mu_0(\mathbf{p})\big)\big(I_1(\mathbf{y}+\mathbf{v}_\mathbf{y})-\mu_1(\mathbf{p}+\mathbf{v}_\mathbf{p})\big)}{\sigma_0(\mathbf{p})\sigma_1(\mathbf{p}+\mathbf{v}_\mathbf{p})} B_{\mathcal{N}}(\mathbf{p}-\mathbf{y}) \, \mathrm{d}\mathbf{y}, \tag{9}$$

with the mean μ and variance σ calculated over the same neighborhood $\mathcal{N}(\mathbf{p})$.The NCC is by construction invariant to linear brightness changes (offset and contrast scaling). In practice, it can be computed efficiently after discretization to the pixel raster, with the help of discrete box filters and integral images. Moreover, truncating the NCC ignores negative correlations [27]: $\mathrm{TNCC}(\mathbf{p},\mathbf{v}) := \min(1, 1-\mathrm{NCC}(\mathbf{p},\mathbf{v}))$. The full data cost is then simply the integral over the image domain: $\mathrm{TNCC}(\mathbf{v}) = \int_{\Omega} \mathrm{TNCC}(\mathbf{p},\mathbf{v}) \, \mathrm{d}\mathbf{p}$.

[1] Some methods thus explicitly enforce the epipolar constraint ("motion stereo"). We refrain from this, as we consider it an instance of stereo matching rather than optical flow estimation.

Mutual information. The mutual information (MI) is a data cost from the alignment literature [25], with even stronger invariance properties. It is popular in stereo matching [13] and has been integrated into optical flow [18]. MI expresses the statistical dependence between two random variables, here image intensities:

$$\mathrm{MI}(I_0, I_1(\mathbf{v})) = \mathrm{H}(I_1(\mathbf{v})) - \mathrm{H}(I_1(\mathbf{v})|I_0) = \mathrm{H}(I_0) + \mathrm{H}(I_1(\mathbf{v})) - \mathrm{H}(I_1(\mathbf{v}), I_0), \quad (10)$$

where $\mathrm{H}(I_1(\mathbf{v}), I_0)$ is the the joint entropy. Mutual information is high if the intensity I_1 can be predicted well from the corresponding I_0; accordingly the negative MI serves as data cost. In practice, intensity statistics are approximated with histograms over pixel values, usually smoothed with an isotropic Gaussian K_ω with kernel size ω.

Census Transform. The original Census transform [28] and its ternary variant [22] have recently found more widespread use, particularly addressing challenging outdoor lighting conditions [17]. This includes methods ranking high in the KITTI benchmark [12, 20]. The (ternary) Census data term at location $\mathbf{p}$ is defined as

$$\mathrm{Cen}(\mathbf{p}, \mathbf{v}) = \textstyle\int_\Omega \mathbb{1}_{c_\epsilon(I_0,\mathbf{p},\mathbf{y}) \neq c_\epsilon(I_1,\mathbf{p}+\mathbf{v},\mathbf{y}+\mathbf{v})} B_{\mathcal{N}}(\mathbf{p}-\mathbf{y})\, \mathrm{d}\mathbf{y} \quad (11)$$

$$\text{with} \quad c_\epsilon(I, \mathbf{p}, \mathbf{q}) = \mathrm{sgn}(I(\mathbf{p}) - I(\mathbf{q}))\mathbb{1}_{|I(\mathbf{p})-I(\mathbf{q})|>\epsilon}, \quad (12)$$

where $\mathbb{1}$ is the indicator function, B is again a box-filter, and $\mathcal{N}(\mathbf{p})$ denotes the corresponding neighborhood. The full Census data cost for a flow field $\mathbf{v}$ is again obtained by integrating over the image domain, $\mathrm{Cen}(\mathbf{v}) = \int_\Omega \mathrm{Cen}(\mathbf{p}, \mathbf{v})\ \mathrm{d}\mathbf{p}$. Although Census has been incorporated in continuous optimization approaches to optical flow, we are not aware of any work that explains how this is done in detail. The Census cost is a piecewise constant function that is neither locally convex nor continuous. Its gradient is 0 or ∞ everywhere, thus there is no obvious linearization (Fig. 1).

To facilitate optimization, we here propose a convex approximation of Eq. (11):

$$\mathrm{CSAD}(\mathbf{p}, \mathbf{v}) = \int_\Omega |I_0(\mathbf{p}) - I_0(\mathbf{y}) - (I_1(\mathbf{p}+\mathbf{v}) - I_1(\mathbf{y}+\mathbf{v}))| B_{\mathcal{N}}(\mathbf{p}-\mathbf{y})\, \mathrm{d}\mathbf{y}, \quad (13)$$

where the "soft" L_1-norm serves as a proxy for the hard thresholding step. We denote the new data term by CSAD, as it is formally a sum of centralized absolute differences. Note that by using absolute differences one foregoes some of the robustness of the ternary census, but gains tractability. As before the full data cost is given by integration over the image domain, *i.e.* $\mathrm{CSAD}(\mathbf{v}) = \int_\Omega \mathrm{CSAD}(\mathbf{p}, \mathbf{v})\ \mathrm{d}\mathbf{p}$. Note also that CSAD bears connections to the widely used gradient constancy assumption (GCA, [6]). Very recently, [10] showed that a continous variant of Census is a generalization of GCA; the discretization bears some resemblance to Eq. 13 and aggregates derivatives of pixel differences.

3.1 Data costs in the primal-dual framework

Several modern optical flow estimation techniques [8, 29], including the primal-dual scheme used here, decouple the optimization of prior and data term, hereby allowing to optimize the data cost independently per pixel. We now describe the update steps (*i.e.*, the proximal operator $(I + \tau\partial G)^{-1}$) for the different data terms, restricted to a single pixel for readability. We first recall the following soft thresholding scheme [15]:

Soft Thresholding. The solution $\bar{x}$ to the optimization problem:

$$\arg\min_{x\in\mathbb{R}} \sum_{i=1}^{n} w_i|x-b_i| + F(x) \tag{14}$$

with $b_i \leq b_{i+1}, \forall i$ and $W_i = -\sum_{j=1}^{i} w_j + \sum_{j=i+1}^{n} w_j, \forall i$, and F being strictly convex and differentiable with a bijective derivative F', can be computed via a median

$$\bar{x} = \text{median}(b_1, \dots, b_n, a_0, \dots a_n), \tag{15}$$

where $a_i = (F')^{-1}(W_i)$. If a data term G can be written as a component-wise weighted sum of L_1-norms as in Eq. (14), one can set $F(x) = x^2/(2\tau)$ and directly solve the proximal map pointwise, through Eq. (15).

BCA. In order to employ Eqs. (4,5), we require a convex data cost G. To that end, we rely on the usual first order Taylor expansion of the brightness of the warped image around an initial solution for the flow field, $\mathbf{v}_0$: $I_1(\mathbf{p}+\mathbf{v}) \approx I_1(\mathbf{p}+\mathbf{v}_0)+(\mathbf{v}-\mathbf{v}_0)^T\nabla I_1$. Our convexified data cost becomes $G_{\text{BCA}}(\mathbf{v}) = \lambda|I_1(\mathbf{p}+\mathbf{v}_0)+(\mathbf{v}-\mathbf{v}_0)^T\nabla I_1 - I_0(\mathbf{p})|$. With that we can write the proximal map for the BCA data cost at pixel $\mathbf{p}$ as

$$(I+\tau\partial G)^{-1}(\hat{\mathbf{v}}) = \arg\min_{\mathbf{v}} \frac{1}{2\tau}(\hat{\mathbf{v}}-\mathbf{v})^2 + \lambda|I_1(\mathbf{p}+\mathbf{v}_0)+(\mathbf{v}-\mathbf{v}_0)^T\nabla I_1 - I_0(\mathbf{p})|. \tag{16}$$

One important observation here is that due to the isotropy of the quadratic term in the proximal map, Eq. (16) can be reduced to a one dimensional problem. In particular, setting $\mathbf{v} = \hat{\mathbf{v}} + \delta\nabla I_1/|\nabla I_1| + \delta^+\nabla^+ I_1/|\nabla^+ I_1|$ the proximal map reduces to:

$$\arg\min_{\delta} \frac{1}{2\tau}\delta^2 + \lambda|\nabla I_1|\left|\underbrace{\frac{I_1(\mathbf{p}+\mathbf{v}_0)+(\hat{\mathbf{v}}-\mathbf{v}_0)^T\nabla I_1 - I_0(\mathbf{p})}{|\nabla I_1|}}_{=:\hat{G}(\hat{\mathbf{v}})} + \delta\right|, \tag{17}$$

hence $\delta^+ = 0$. Here ∇^+ is a vector orthogonal to the gradient. This can also be generalized to different data terms by using brightness linearization of the warped image. Applying Eq. (15) we can solve for the optimal δ and derive a soft-thresholding scheme: $(I+\tau\partial G)^{-1}(\hat{\mathbf{v}}) := \hat{\mathbf{v}} + \nabla I_1/|\nabla I_1| \cdot \text{median}\{-\hat{G}(\hat{\mathbf{v}}), \pm\lambda\tau|\nabla I_1|\}$.

NCC. The (T)NCC cost function is not convex, and a closed form solution for the proximal map does not exist. Following [27], a second order Taylor expansion can be used to build a convex approximation; off-diagonal entries of the Hessian are dropped to make it positive definite. At pixel $\mathbf{p}$ the convexified data term G_{TNCC} becomes

$$G_{\text{TNCC}}(\mathbf{v}) = \lambda\big(\text{TNCC}(\mathbf{p},\mathbf{v}_0)+(\mathbf{v}-\mathbf{v}_0)^T\nabla_{\text{TNCC}}+\frac{1}{2}(\mathbf{v}-\mathbf{v}_0)^T\nabla^2_{+,\text{TNCC}}(\mathbf{v}-\mathbf{v}_0)\big), \tag{18}$$

where the Taylor expansion was developed at $\mathbf{v}_0 = (u_0, v_0)$, and the modified Hessian is given by $\nabla^2_{+,\text{TNCC}} = \text{diag}(\max(0, \partial^2\text{TNCC}/\partial^2 u), \max(0, \partial^2\text{TNCC}/\partial^2 v))$.

Mutual information. The local convexification of the MI data cost from [18] is similar to the second order approximation introduced in Eq. (18), but here applied to the MI data cost $G_{\text{MI}} = -\lambda\text{MI}$. We again modify the second order term to ensure positive definiteness, and reduce the problem to 1D (Eq. 17). For details please refer to [18].

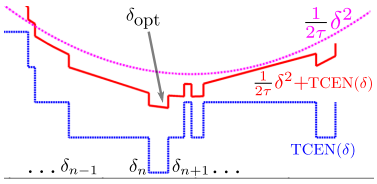

Fig. 1. Optimization of the ternary census data cost: The proximal map is the sum of the piecewise constant census score and a quadratic.

CSAD. Similar to the BCA, we linearize the intensity around the center pixel, which allows reducing the problem to 1D (Eq. 17). The proximal map at $\mathbf{p}$ again takes the form of Eq. (14), which can be solved using the soft-thresholding scheme. Due to the constant weights, the solution to the proximal map reduces to an efficient median search. In preliminary experiments we also considered the generalization of GCA described in [10], by introducing appropriate spatial weights. The weighting did not improve the results, thus we did not pursue it further. Moreover, we note that we find CSAD to work best with larger windows (see below), suggesting a benefit over GCA.

Census. Rather than use the CSAD approximation to the ternary census, we now show how to solve the proximal map directly for the census data cost. We start by linearizing the brightness in Eq. (11) (at the patch center). Following the same reasoning as for BCA (*cf*. Eq. 17), the optimal displacement must be in the direction of the image gradient. Hence, we again need to solve a one-dimensional optimization problem and use $I_1(\mathbf{p}+\mathbf{v}_0)+\delta\nabla I_1/|\nabla I_1|$ as replacement for the intensity at the center pixel in Eq. (11); the intensity is now a function of δ. By inspection of that function (Fig. 1), we can identify discontinuities, *i.e*. locations where one summand of the census cost changes (or can change). The trick is to precompute these at most $2|\mathcal{N}(\mathbf{p})|$ locations $\{\delta_n\}$, which we sort in increasing order. At each δ_n we determine whether the cost increases or decreases by 1, or stays constant. We can then determine the value of the census cost using a cumulative sum of the cost changes. This allows us to efficiently find the minimum of the proximal map by considering the identified candidate locations, and taking into account the quadratic penalty $\frac{1}{2\tau}\delta^2$. The data cost changes depend only on I_0 and can be computed once per pyramid level of the coarse-to-fine scheme; the respective $\{\delta_n\}$ need to be computed and sorted only once per warp.

4 Evaluation

We evaluate the different data costs on two datasets, the well established Middlebury data set [1], containing 8 test images with ground truth, and the more recent KITTI dataset [9] with 194 scenes. See Sec. 2 for a discussion. Parameters have been determined empirically, for best performance on the KITTI training set. In all experiments we apply coarse-to-fine estimation with a pyramid scale factor of 0.9, 40 warps and 5 inner iterations per pyramid level, and outlier removal through median filtering after each pyramid level [26]. Derivatives are computed with bicubic interpolation [24]. For our evaluation on KITTI we use TGV^2_α regularization with $\alpha_0 = 5, \alpha_1 = 1$. In case of Middlebury, we use TV instead. The weight for the data cost is set to: $\lambda = 8/9$ for TNCC, $\lambda = 80/|\mathcal{N}-1|$ for CSAD and Census, and to $\lambda = 25$ for BCA. The threshold for Census is set to $\epsilon = 0.005$. For MI we use a 15×15 Gaussian filter with $\omega = 3$.

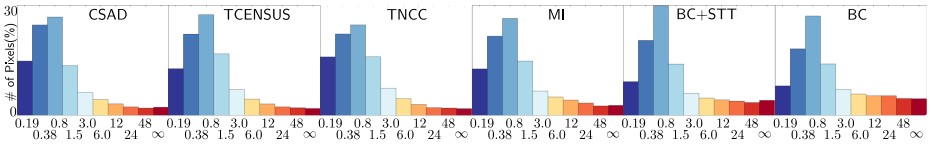

Fig. 2. Histograms of endpoint errors for different data costs, on complete KITTI training set. Methods are ordered w.r.t. the number of pixels with endpoint error < 1.5px. See text for details.

Middlebury. Table 1 reports the average endpoint error (AEP) for the Middlebury *training* set. The accuracy difference of the pixel-based data terms appears to be rather small: MI is on par with simple BCA; preprocessing with STT does not appear to help at all, performing worse than BCA in every case. Among the patch-based data terms CSAD achieved the best results. Compared to Census, soft rather than hard thresholding of brightness differences allows for a more fine-grained localization, see Fig. 2. Median filtering visibly increases the smoothness of the Census flow vectors especially for the smallest patch size. In general, larger patch sizes appear to perform better than smaller ones. The most important finding for Middlebury is that pixel-based data costs perform as well as patch-based ones.[2] We attribute this to the controlled lighting conditions. Difficulties lie rather in the non-rigid motion patterns, occlusions, or repetitive texture. Illumination invariant patch-based data costs cannot improve the results under these conditions. Importantly, however, the results do not deteriorate either.

KITTI. On the full KITTI training set in contrast (Tab. 2, left), we observe a clear performance improvement of the patch-based data costs over the pixel-based ones. Among the pixel-based ones, MI clearly outperforms BCA (with and without STT preprocessing). For the patch-based data terms, every measure is best for at least two settings, thus showing no clear winner. However, Census at the largest patch size can be identified as the overall winner. Interestingly, the performance of NCC peaks at small window sizes, while for CSAD and Census a larger patch size works best. However, the performance gain w.r.t. patch size saturates at 7×7 for all patch-based costs.

Fig. 2 shows the endpoint error distribution for all data costs, ordered by decreasing number of inliers (< 1.5px). CSAD and NCC show a higher fraction of flow vectors with low EPE (< 0.38), *i.e.* they offer a higher localization accuracy than the other data costs. We also evaluated one data cost (CSAD, 5×5) on the official test portion

[2] We note that patch-based data costs are challenged by rotational motion, however these are not very prominent here (nor in many other application scenarios).

data cost	win.size	Dimetrodon	Grove2	Grove3	Hydrangea	Rubberwhale	Urban2	Urban3	Venus
BC	—	0.19	0.21	0.64	0.21	0.15	0.35	0.69	0.34
BC+STT	—	0.22	0.22	0.67	0.21	0.15	0.39	0.85	0.39
MI	—	**0.17**	0.21	0.65	0.20	**0.12**	0.35	0.73	0.33
CSAD	3×3	0.24	0.25	0.66	**0.18**	0.14	0.37	0.79	0.34
	5×5	0.19	0.23	0.64	**0.18**	**0.12**	0.36	0.68	0.33
	7×7	0.18	0.23	**0.61**	**0.18**	**0.12**	0.36	0.63	**0.32**
NCC	3×3	0.21	0.22	0.74	**0.18**	**0.12**	0.36	0.63	0.34
	5×5	0.18	**0.20**	0.70	0.19	0.13	**0.34**	0.61	0.34
	7×7	**0.17**	0.21	0.72	0.20	0.15	0.35	0.69	0.34
Census	3×3	0.31	0.24	0.72	0.22	0.17	0.45	0.74	0.40
	5×5	0.25	0.22	0.68	0.20	0.14	0.41	0.59	0.38
	7×7	0.24	0.21	0.66	0.19	0.14	0.39	**0.56**	0.36

Table 1. Average endpoint error [px] on *Middlebury* training dataset.

		full KITTI training set										Illumination Changes									
		KITTI metric								**AEP**		**KITTI metric**					**AEP**				
occluded pix.		×				✓				×	✓	✓					✓				
		2px	3px	4px	5px	2px	3px	4px	5px			#44	#11	#15	#74	Av.	#44	#11	#15	#74	Av.
BC	—	20.7	17.2	15.0	13.4	28.9	25.1	22.6	20.7	3.9	7.7	45.6	35.6	67.1	90.6	59.7	11.2	9.7	21.0	45.7	21.9
BC+STT	—	14.3	11.5	10.0	8.8	23.0	19.8	17.9	16.3	2.8	6.7	28.7	29.0	40.7	64.2	40.7	9.3	9.0	8.4	23.2	12.5
MI	—	12.2	9.4	7.9	6.9	20.9	17.5	15.5	13.8	2.5	5.3	22.5	33.8	19.9	62.0	34.6	6.3	10.2	5.7	**20.1**	10.6
CSAD	3×3	9.9	7.5	6.3	5.3	17.2	13.9	12.1	10.7	1.8	4.3	24.0	24.4	15.6	57.6	30.3	11.8	7.4	5.0	22.8	11.8
	5×5	**9.6**	7.2	6.0	5.1	17.0	13.8	11.9	10.6	**1.7**	4.2	22.1	**23.3**	16.4	59.8	30.4	10.2	6.6	**3.7**	23.3	10.9
	7×7	9.7	7.3	6.0	5.1	17.1	13.9	12.1	10.7	**1.7**	4.3	19.7	24.2	17.7	**58.6**	30.1	7.4	7.2	3.9	23.0	**10.4**
NCC	3×3	10.3	7.3	5.9	5.0	17.2	13.5	**11.4**	**10.0**	1.8	3.8	18.1	33.6	**14.2**	60.5	31.6	**6.1**	13.6	4.9	23.4	12.0
	5×5	10.1	7.3	6.0	5.1	17.2	13.6	11.6	10.2	1.8	**3.7**	21.3	32.9	15.5	59.3	32.3	9.3	13.6	4.4	23.3	12.7
	7×7	10.7	7.9	6.5	5.5	17.8	14.3	12.2	10.8	1.9	4.1	**17.9**	32.9	16.9	58.7	31.6	6.2	13.2	4.2	23.4	11.8
Census	3×3	10.4	7.7	6.4	5.5	17.7	14.2	12.3	10.9	2.0	4.5	19.2	27.8	18.6	59.9	31.4	6.4	9.0	6.6	23.9	11.5
	5×5	9.7	7.0	5.7	4.8	17.0	13.5	11.6	10.2	**1.7**	4.0	19.0	25.3	17.1	59.0	30.1	6.6	7.3	6.0	23.6	10.9
	7×7	**9.6**	**6.9**	**5.6**	**4.7**	**16.9**	**13.3**	**11.4**	**10.0**	**1.7**	**3.7**	18.7	24.4	15.9	59.5	**29.6**	6.3	**6.3**	5.0	23.7	**10.4**

Table 2. *KITTI* metric (percentage of flow vectors above 2/3/4/5 pixels of endpoint error) and average endpoint error [px], for the complete *KITTI* training set, as well as for the illumination images selected for the GCPR special session.

of the KITTI benchmark, where our method ("Data-Flow") is ranked 6^{th} at the time of publication, which shows that our testbed is state-of-the-art. We observe 8.2% outliers excluding and 15.8% outliers including occluded regions.

Illumination changes. We also report results for particularly challenging test images as part of the *Robust Optical Flow Challenge* (Tab. 2, right). We did not adapt the parameters in any way. While the outlier percentages are generally high, owing to the difficulty of the challenge, the patch-based data costs allow the approach to significantly outperform the official baseline techniques. On average 7×7 Census and 7×7 CSAD perform slightly better than the remaining patch-based data costs. The gap to the pixel-based error metrics is again large, with the exception of MI, which produces only 5% more outliers than the best patch-based measure. On the selected subset our TNCC implementation has about 20% fewer outliers than [27], which uses the same data term, but TV regularization. This might be an indication that the data term alone cannot compensate for using a prior that is not suitable for the scenario.

5 Conclusion

Based on a state-of-the-art testbed and challenging image data, we provided an evaluation of several pixel-based and patch-based data costs. While on the standard Middlebury dataset, patch-based measures cannot provide a clear benefit, they show significant gains on the more challenging KITTI dataset. Overall, the Census transform and the proposed CSAD variant, which is well-suited for standard continuous optimization, show a slight overall performance lead. By avoiding thresholding, CSAD showed to be particularly well suited for accurate flow localization.

References

1. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. In: ICCV (2007)
2. Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. IJCV (1994)

3. Black, M.J., Anandan, P.: A framework for the robust estimation of optical flow. In: ICCV (1993)
4. Bleyer, M., Rhemann, C., Rother, C.: PatchMatch Stereo – Stereo matching with slanted support windows. In: BMVC (2011)
5. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sciences 3(3), 492–526 (2010)
6. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: ECCV (2004)
7. Bruhn, A., Weickert, J., Kohlberger, T., Schnörr, C.: A multigrid platform for real-time motion computation with discontinuity-preserving variational methods. IJCV 70(3) (2006)
8. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. JMIV 40(1), 120–145 (2011)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR (2012)
10. Hafner, D., Demetz, O., Weickert, J.: Why is the census transform good for robust optic flow computation? In: SSVM (2013)
11. Haussecker, H.W., Fleet, D.J.: Computing optical flow with physical models of brightness variation. In: CVPR. vol. 2 (2000)
12. Hermann, S., Klette, R.: Hierarchical scan line dynamic programming for optical flow using semi-global matching. In: ACCV workshop on Intelligent Mobile Vision (2012)
13. Hirschmüller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: CVPR (2007)
14. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. 17(1-3) (1981)
15. Li, Y., Osher, S.: A new median formula with applications to PDE based denoising. Commun. Math. Sci. 7(3), 741–753 (2009)
16. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI (1981)
17. Müller, T., Rabe, C., Rannacher, J., Franke, U., Mester, R.: Illumination-robust dense optical flow using census signatures. In: DAGM (2011)
18. Panin, G.: Mutual information for multi-modal, discontinuity-preserving image registration. In: ISVC (2012)
19. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: ICCV (2011)
20. Ranftl, R., Gehrig, S., Pock, T., Bischof, H.: Pushing the limits of stereo using variational stereo estimation. In: Intelligent Vehicle Symposium (2012)
21. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1-4), 259–268 (1992)
22. Stein, F.: Efficient computation of optical flow using the census transform. In: DAGM (2004)
23. Steinbrücker, F., Pock, T., Cremers, D.: Advanced data terms for variational optic flow estimation. In: VMV (2009)
24. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: CVPR (2010)
25. Viola, P., Wells III, W.: Alignment by maximization of mutual information. In: ICCV (1995)
26. Wedel, A., Pock, T., Zach, C., Cremers, D., Bischof, H.: An improved algorithm for TV-L1 optical flow. In: Proc. of the Dagstuhl Motion Workshop. LNCS, Springer (2008)
27. Werlberger, M., Pock, T., Bischof, H.: Motion estimation with non-local total variation regularization. In: CVPR (2010)
28. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: ECCV (1994)
29. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. In: DAGM (2007)