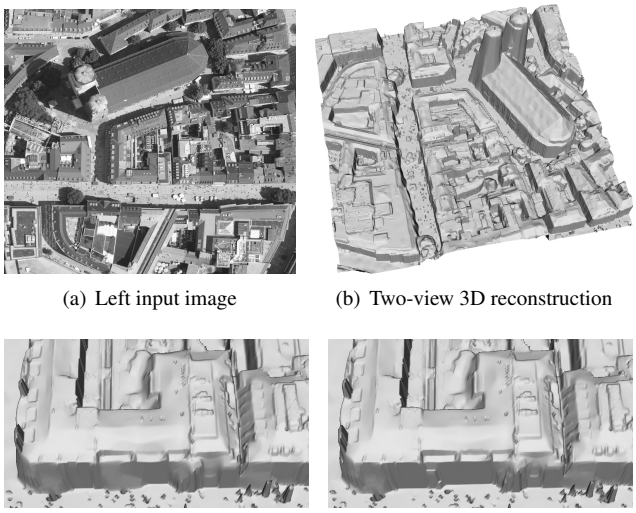
Fast and Accurate Large-scale Stereo Reconstruction using Variational Methods

Georg Kuschk[1,2] Daniel Cremers[1]
[1] TU Munich, Germany
[2] German Aerospace Center (DLR), Germany

Abstract

This paper presents a fast algorithm for high-accuracy large-scale outdoor dense stereo reconstruction of man-made environments. To this end, we propose a structure-adaptive second-order Total Generalized Variation (TGV) regularization which facilitates the emergence of planar structures by enhancing the discontinuities along building facades. As data term we use cost functions which are robust to illumination changes arising in real world scenarios. Instead of solving the arising optimization problem by a coarse-to-fine approach, we propose a quadratic relaxation approach which is solved by an augmented Lagrangian method. This technique allows for capturing large displacements and fine structures simultaneously. Experiments show that the proposed augmented Lagrangian formulation leads to a speedup by about a factor of 2. The brightness-adaptive second-order regularization produces sub-disparity accurate and piecewise planar solutions, favoring not only fronto-parallel, but also slanted planes aligned with brightness edges in the resulting disparity maps. The algorithm is evaluated and shown to produce consistently good results for various data sets (close range indoor, ground based outdoor, aerial imagery).



(a) Left input image (b) Two-view 3D reconstruction

(c) Zoom-in: Reconstruction using TGV and an anisotropic diffusion tensor based on pixelwise gradients (d) Zoom-in: Improvements along discontinuities by additionally using high-level edge information

Figure 1. Detailled stereo reconstruction using two 1000×1000 wide-baseline aerial images, taking 10 seconds on common GPUs.

1. Introduction

In the past few years, Total Variation based methods for minimizing energy functionals arising in common computer vision problems have been given a lot of attention in the research community. These algorithms are very well-suited for parallelization and, together with the recent advances of GPU-based computational power, lead to efficient algorithms, solving these optimization problems globally optimal. Recently published work solving e.g. the optical flow or stereo estimation problem can be found in [16], [14], [9], [11]. Total Generalized Variation (TGV) was originally introduced in [2] as a higher-order extension of Total Variation minimization (TV) and favors the solution to consist of piecewise polynomial functions (e.g. fronto-parallel, affine, quadratic). Like the original TV formulation, the TGV regularizer also is convex and allows for computation of the global optimum. In the following two years, the second-order variant of TGV has been applied to depth map fusion in [10] and dense stereo estimation in [11], basically assuming that the surface to reconstruct is locally planar and not implying fronto-parallel constraints only. For being able to use robust cost functions which are usually highly non-linear, a typical choice is to linearize the costs inside a coarse-to-fine strategy (see e.g. [11]). The main drawback of this approach is that fine scene-details which are not captured in the lower pyramid levels are highly likely to be missing completely in the final reconstruction. Applying TGV as regularizer for stereo estimation, the energy functional we will use throughout the rest of the paper and need to minimize reads

$$E = \int_{\Omega} \{\lambda_s |G(\nabla \mathbf{u} - \mathbf{v})| + \lambda_a |\nabla \mathbf{v}| + \lambda_d C(\mathbf{u})\} \, \mathrm{d}\mathbf{x} \quad (1)$$

with $\mathbf{u}(\mathbf{x}) \in \Gamma$ the disparity/depth map to solve for (Γ being the disparity search space), an additional vector field $\mathbf{v}$ and Ω being the image space $\mathbb{R}^{M \times N}$. Note that for brevity, we just write $\mathbf{u}, \mathbf{v}$ instead of $\mathbf{u}(\mathbf{x}), \mathbf{v}(\mathbf{x})$. So instead of just enforcing the norm of the gradient of $\mathbf{u}$ to be minimal, which equals favoring fronto-parallel surfaces, the additional vector field $\mathbf{v}$ gets subtracted from the gradient of $\mathbf{u}$ and in turn is also forced to have low variation. Therefore, piecewise affine functions are being favored, as these have a constant gradient whose derivative tends to zero. The values $\lambda_s, \lambda_a, \lambda_d$ are scalar weights and balance the impact of the smoothness term, the affine term and the data term.

The linear operator G in Equation 1 serves to adapt the amount of regularization locally, depending on some information derived from the input images. A famous choice for G is for example the anisotropic Nagel-Enkelmann operator [7], which, in addition to the original paper, has been widely used and modified throughout the literature ([16], [11]). However, all these methods have in common, that they compute an adaptive regularization weight based on the local image gradient at the considered pixel solely. This usually improves the sharpness along discontinuities, but does not necessarily impose straight edges along man made structures. To improve the accuracy of the stereo estimation along these straight-line discontinuities, we integrate an adaptive regularization weight based on detected high-level line segments, which is inherently easy to integrate into the proposed global optimization framework.

Unfortunately, we cannot solve Equation 1 directly with e.g. a primal-dual gradient based approach [9], since the data term should be a strong and reliable cost function to fit our needs of being robust against some amount of change in perspective and illumination (and therefore in general non-convex). This problem often is bypassed by linearizing the cost function and solving the resulting convex problem. Since this 1st order Taylor approximation of the cost function is only valid locally, the whole algorithm needs to be wrapped into a coarse-to-fine warping framework [3], which we explicitly want to avoid to not loose fine structures already in the coarsest level. In the following section, we will explain our solution to this minimization problem.

2. Edge-segment based adaptive regularization

The anisotropic diffusion tensor G in Equation 1 serves the purpose of an anisotropic weighting of the regularizer based on the image gradient. It enforces low regularization/smoothness along image edges, and high smoothness in homogenous image regions. It is based on the Nagel-Enkelmann operator [7] and was proposed in [16]:

$$G = \exp(-a \cdot |\nabla I_{ref}|^b) \cdot nn^T \; + \; n^\perp n^{\perp T} \qquad (2)$$

with the direction of the image gradient $n = \binom{n_x}{n_y} = \frac{\nabla I_{ref}}{|\nabla I_{ref}|}$, an perpendicular vector $n^\perp$ and weighting parameters a, b.

However, as this diffusion tensor is based on pixelwise gradients (incorporating spatial context to a minor degree by a prior Gaussian convolution), it does not provide a strong and consistent regularization direction for small low-contrast edges as shown in Figure 2.

Using high-level edge segments as additional a priori information is a logical choice for guiding the optimization framework to straight-line discontinuity reconstructions. However, the main problem with this approach is the robustness of the edge detection, as for most edge detection algorithms (e.g. Canny [4]), textured regions result in a high edge density and therefore many false detections. A second problem for heterogenous image data is the need to manually tune the parameters for each group of images separately, to obtain reasonable results.

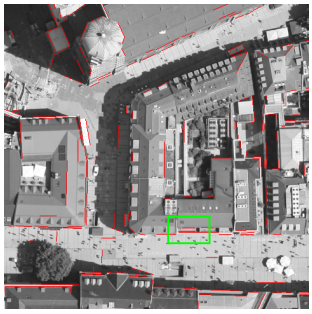
The recently introduced Fast Line Segment Detector (LSD) [15] adresses both of these problems and gives outstanding results while being computationally quite efficient. The integration of the edge-segments into the optimization framework is straight forward, as we repeat the process described in Equation 2 with the Gauss-convoluted binary mask of detected edge segments as input image, resulting in a second diffusion tensor G'. We obtain the combined diffusion tensor by updating the values of G with the values of G' at the position of detected lines (see Figure 2).

3. Fast optimization by quadratic splitting and augmented Lagrangian

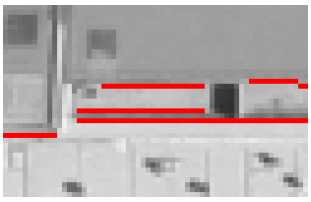
In [14], a quadratic relaxation between the convex regularizer and the non-convex data term was proposed for minimizing a Total Variation based optical flow energy functional and [8] used a similar approach for image driven and TV-based stereo estimation. We build upon these ideas and split the image driven TGV stereo problem from Equation 1 into two subproblems and, using quadratic relaxation, couple the convex regularizer $R(\mathbf{u})$ and non-convex data term $C(\mathbf{u})$ through an auxiliary variable $\mathbf{a}$:

$$E = \int_\Omega R(\mathbf{u}) + C(\mathbf{a}) + \frac{1}{2\theta}(\mathbf{u} - \mathbf{a})^2 \, \mathrm{d}\mathbf{x} \; . \qquad (3)$$

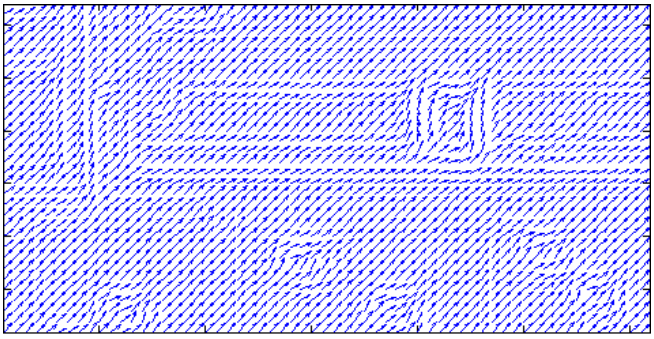
By iteratively decreasing $\theta \to 0$, the two variables $\mathbf{u}, \mathbf{a}$ are drawn together, enforcing the equality constraint $\mathbf{u} = \mathbf{a}$. As an alternative, we incorporate this equality constraint not uniformly for each pixel, but via an additional augmented Lagrange multiplier $\mathbf{L}$ (see e.g. [1]) and optimize for it as well. The resulting energy minimization problem based on



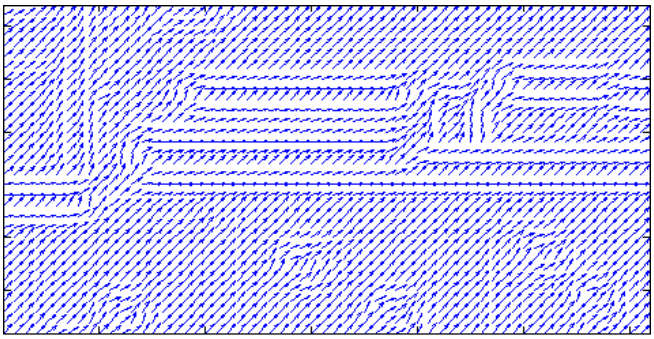
(a) Part of the left stereo image, overlayed with detected edges



(b) Zoom-in area of green rectangle



(c) Zoom-in area: Vector field of the anisotropic diffusion tensor based on pixelwise gradients



(d) Vector field additionally incorporating high-level edge information

Figure 2. Influence of additional high-level edge priors on the anisotropic regularization: Due to low contrast, the Nagel-Enkelmann operator in c) cannot capture the building edge of b) very well. Using additional edge information d) improves the regularization direction.

Equation 3 then reads as follows

$$\mathbf{u} = \underset{\mathbf{u}}{\operatorname{argmin}} \{ \lambda_s |G(\nabla \mathbf{u} - \mathbf{v})| + \lambda_a |\nabla \mathbf{v}| + \lambda_d C(\mathbf{a}) + \mathbf{L}(\mathbf{u} - \mathbf{a}) + \frac{1}{2\theta}(\mathbf{u} - \mathbf{a})^2 \} \quad (4)$$

Our experiments showed that this improves the robustness of the algorithm w.r.t. the choice of the θ-sequence and additionally speeds up the algorithm by a factor of 2 (see Figure 4).

While the regularization term is convex in $\mathbf{u}$ and can be solved efficiently using a primal-dual approach for a fixed auxiliary variable $\mathbf{a}$, the non-convex data term can be solved point-wise by an exhaustive search over a set of discretely sampled disparity values. This process is done alternatingly in an iterative way.

3.1. Convex solution

To solve for the disparity map $\mathbf{u} \in \mathbb{R}^{M \times N}$ (in the following written as stacked vector $\mathbb{R}^{MN \times 1}$) in the regularizer term of Equation 4, we need to overcome the non-differentiable L_1-norm, which complicates any gradient descent based minimization scheme. To this end we apply the Legendre-Fenchel transform to obtain the dual formulation / conjugate of our L_1 regularizers

$$\lambda \|AG\mathbf{u}\|_1 = \underset{\|\mathbf{p}\| \leq \lambda}{\operatorname{argmax}} \{ \langle AG\mathbf{u}, \mathbf{p} \rangle \} \quad (5)$$

where the matrix multiplication $A\mathbf{u}$ computes the $2MN \times 1$ gradient vector and $G \in \mathbb{R}^{M \times N}$ contains the element-wise weighting factors. Applied to our problem, we obtain the conjugates

$$\lambda_s \cdot \|G(\nabla \mathbf{u} - \mathbf{v})\|_1 = \max_{\mathbf{p} \in P} \{ \langle G(\nabla \mathbf{u} - \mathbf{v}), \mathbf{p} \rangle \} \quad (6)$$
$$\lambda_a \cdot \|\nabla \mathbf{v}\|_1 = \max_{\mathbf{q} \in Q} \{ \langle \nabla \mathbf{v}, \mathbf{q} \rangle \}$$

such that the saddle-point problem in the primal variables $\mathbf{u}, \mathbf{v}$ and their dual correspondences $\mathbf{p}, \mathbf{q}$ with constraints $P = \{\mathbf{p} \in \mathbb{R}^{2MN} : \|\mathbf{p}\|_\infty \leq \lambda_s\}$ and $Q = \{\mathbf{q} \in \mathbb{R}^{4MN} : \|\mathbf{q}\|_\infty \leq \lambda_a\}$, coupled with the data term is $\max_{\mathbf{p},\mathbf{q}} \min_{\mathbf{u},\mathbf{v},\mathbf{a}} \{E\}$ with

$$E = \langle G(\nabla \mathbf{u} - \mathbf{v}), \mathbf{p} \rangle + \langle \nabla \mathbf{v}, \mathbf{q} \rangle + \lambda_d C(\mathbf{a}) + \mathbf{L}(\mathbf{u} - \mathbf{a}) + \frac{1}{2\theta}(\mathbf{u} - \mathbf{a})^2 \quad (7)$$

Fixing the variables $\mathbf{a}$ and $\mathbf{L}$, we obtain the minimum of Equation 7 for $\partial_{\mathbf{u},\mathbf{v},\mathbf{p},\mathbf{q}} E(\mathbf{u}, \mathbf{v}, \mathbf{a}, \mathbf{p}, \mathbf{q}) = 0$ and using an iterative gradient descent in the primal variables and gradient ascent in the dual variables.

3.2. Non-convex solution

To solve for the auxiliary variable $\mathbf{a}$ in the data term of Equation 4, we keep the variables $\mathbf{u}, \mathbf{L}$ fixed and perform a point-wise exhaustive search over all $\mathbf{a}(\mathbf{x}) \in \Gamma$

$$\min_{\mathbf{a}(\mathbf{x}) \in \Gamma} \left\{ \lambda_d C(\mathbf{a}) + \mathbf{L}(\mathbf{u} - \mathbf{a}) + \frac{1}{2\theta}(\mathbf{u} - \mathbf{a})^2 \right\} \quad (8)$$

Note that in order to retain the TGV smoothness, it is necessary to perform the exhaustive search using subdisparity sampling steps. As this may look computational expensive at first glance, it does not affect the overall performance in a measurable way if implemented with care (see Section 4).

3.3. Augmented Lagrangian update

According to e.g. [1], the Lagrange multiplier L is updated by $\mathbf{L}^{n+1} = \mathbf{L}^n + \frac{1}{2\theta^n}(\mathbf{u} - \mathbf{a})$, with the augmented penalty function $\frac{1}{2\theta^n}$ monotonically increasing as $\theta^n \to 0$.

4. Algorithm

In this section we will describe how to solve the energy minimization problem stated in Equation 4. As a first step, since the stereo estimation should work with various scales in depth and different cost functions as well without having to adjust the parameters for each dataset, we initially norm both $\mathbf{u} \to [0, 1]$ and the costs $C \to [0, 1]$. Doing so, we can fix nearly all parameters internally and only need to expose the weighting factors λ_d, λ_s, balancing the impact of the data term and smoothness term, to be set by the user. After evaluating the algorithm for a variety of scenarios (indoor, ground-based outdoor, aerial) and benchmarks (see Section 5), we obtained the best results for $\lambda_a = 8\lambda_s$ and fix this value to not bother the user with the weighted impact of the affine term additionally. The complete optimization of the proposed energy functional in Equation 4 is done iteratively, initializing the primal variable with the disparity value associated to the data cost minimum (winner-takes-all solution), $\mathbf{u}^0 = \mathbf{a}^0 = \text{argmin}_{\mathbf{a}(\mathbf{x})\in\Gamma} C(\mathbf{x}, \mathbf{a}(\mathbf{x}))$, setting the dual variables to zero ($\mathbf{p}^0 = 0$, $\mathbf{q}^0 = 0$), and starting with iteration $n = 0$ and $\theta^0 = 1$.

1. Fixing $\mathbf{a}^n$ and $\mathbf{L}^n$, run the primal-dual optimization for a number of inner iterations, performing gradient ascents on the dual variables $\mathbf{p}, \mathbf{q}$ and gradient descents on the primal variables $\mathbf{u}, \mathbf{v}$:
   for $i = 1 : nIterSmooth$ do

$$\mathbf{p}^{n+1} = \Pi_P\left(\mathbf{p}^n + \tau_p G(\nabla\hat{\mathbf{u}}^n - \hat{\mathbf{v}}^n)\right)$$
$$\mathbf{q}^{n+1} = \Pi_Q\left(\mathbf{q}^n + \tau_q \nabla\hat{\mathbf{v}}^n\right)$$
$$\mathbf{u}^{n+1} = \Pi_U\left(\frac{\mathbf{u}^n + \tau_u \text{div}\left(G\mathbf{p}^{n+1}\right) - \tau_u\mathbf{L}^n + \frac{\tau_u}{\theta^n}\mathbf{a}^n}{1 + \frac{\tau_u}{\theta^n}}\right)$$
$$\mathbf{v}^{n+1} = \mathbf{v}^n + \tau_v(\mathbf{p}^{n+1} + \text{div}\mathbf{q}^{n+1})$$
$$\hat{\mathbf{u}}^{n+1} = 2\mathbf{u}^{n+1} - \mathbf{u}^n$$
$$\hat{\mathbf{v}}^{n+1} = 2\mathbf{v}^{n+1} - \mathbf{v}^n$$

2. Fixing $\mathbf{u}^{n+1} = \tilde{\mathbf{u}}$, perform a point-wise search

$$\mathbf{a}^{n+1} = \underset{\mathbf{a}(\mathbf{x})\in\Gamma}{\text{argmin}}\left\{\lambda_d C(\mathbf{a}) + \mathbf{L}^n(\tilde{\mathbf{u}} - \mathbf{a}) + \frac{(\tilde{\mathbf{u}} - \mathbf{a})^2}{2\theta^n}\right\}$$

3. Update $\mathbf{L}^{n+1} = \mathbf{L}^n + \frac{1}{2\theta^n}(\mathbf{u}^{n+1} - \mathbf{a}^{n+1})$

4. If $n < n_{stop}$, update $\theta^{n+1} = \theta^n(1 - \beta n)$, $n = n + 1$, goto step (1) else stop

To ensure that $\|\mathbf{p}\|_\infty \leq \lambda_s$ and $\|\mathbf{q}\|_\infty \leq \lambda_a$, the proximal mappings above are given as $\Pi_P(\mathbf{p}) = \frac{\mathbf{p}}{\max\{1, \|\mathbf{p}\|/\lambda_s\}}$ and $\Pi_Q(\mathbf{q}) = \frac{\mathbf{q}}{\max\{1, \|\mathbf{q}\|/\lambda_a\}}$ and for keeping $\mathbf{u}$ in valid range, we use Π_U as the truncation of $\mathbf{u}^{n+1}$ onto the interval $[0, 1]$. Also note, that in the analytical derivation of the primal-dual scheme above, we require the gradient and divergence operators to be negative adjoint, such that $\langle\nabla\mathbf{u}, \mathbf{p}\rangle = -\langle\mathbf{u}, \text{div}\mathbf{p}\rangle$ and $\langle\nabla\mathbf{v}, \mathbf{q}\rangle = -\langle\mathbf{v}, \text{div}\mathbf{q}\rangle$. Therefore we use finite forward differences with Neumann boundary conditions for the gradient operators and for the divergence operators finite backward difference with Dirichlet boundary conditions. The step sizes of the gradient ascents/descents are bound to the norm of the gradient/divergence operators and are set to $\tau_u = \tau_p = 1/\sqrt{12}$ and $\tau_v = \tau_q = 1/\sqrt{8}$, as detailed in [5]. The parameter β controls how fast the convex and non-convex solution are drawn together (by decreasing θ) and is fixed to $\beta = 10^{-3}$, while the whole algorithm stops, if $n > 80$. For the number of primal-dual iterations, we set $nIterSmooth = 150$.

As already mentioned in Section 3, retaining the subdisparity smoothness resulting from the continuous TGV solution requires subdisparity accurate results of the exhaustive search as well. Therefore, after obtaining an integer solution for the disparity $\mathbf{a}$ which minimizes the energy

$$\underset{\mathbf{a}}{\text{argmin}}\left\{\lambda_d C(\mathbf{a}) + \mathbf{L}(\mathbf{u} - \mathbf{a}) + \frac{1}{2\theta}(\mathbf{u} - \mathbf{a})^2\right\}, \quad (9)$$

we compute the subdisparity solution as the minimum of a parabola, fitted through the obtained integer minimum and its adjacent values at ± 1 disparities (see Figure 3). Parametrizing the parabola as $C(\mathbf{a} + t) = at^2 + bt + c$, the coefficients are computed using the abovementioned 3 datapoints and corresponding $t \in \{-1, 0, 1\}$. Substituting $\tilde{\mathbf{a}} = \mathbf{a} + t$, $C(\tilde{\mathbf{a}}) = at^2 + bt + c$ and optimizing for the parameter t, we obtain the subdisparity refinement $\tilde{t} \in [-\frac{1}{m}, \frac{1}{m}]$ as

$$\tilde{t} = \frac{\frac{\mathbf{u}-\mathbf{a}}{\theta m} - \lambda b - \frac{\mathbf{L}}{m}}{\left(2\lambda a + \frac{1}{\theta m^2}\right)}, \quad (10)$$

with $m = |\Gamma|$ being the number of disparities.

Finally, due to its iterative and locally confined computations per iteration, the algorithm is very well-suited for parallelization and therefore implemented on GPU.

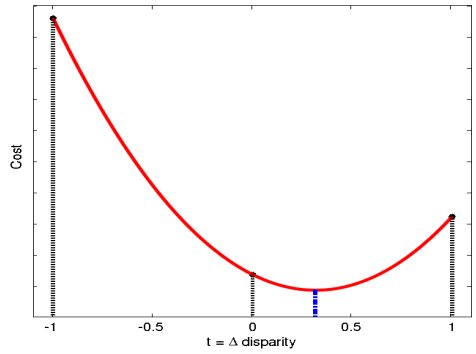


Figure 3. Subdisparity accurate results are required in the exhaustive search step, to retain the continuous solution of the prior TGV step

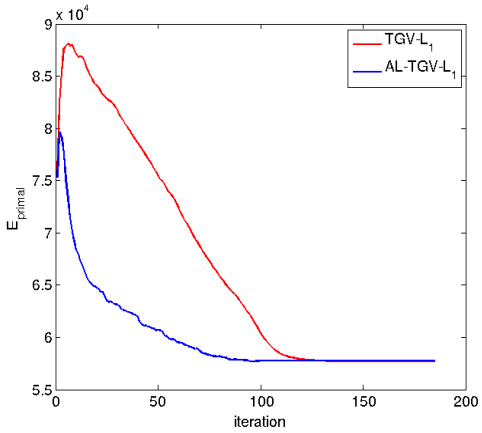


Figure 4. Evolution of the primal energy of Equation 3, with and without augmented Lagrangian. The runtime is dominated by the primal-dual algorithm, such that the additional Lagrange multiplier $\mathbf{L}$ has a neglectable influence and the runtime per iteration is basically the same for the two algorithms.

5. Evaluation

Our algorithm is evaluated on three different data sets and in case RGB images are available, only the gray image will be used. If more than two views are available, only two of them will be used, in order to demonstrate our algorithm on two-view stereo scenarios. For all datasets, we used the Census transform [18] with windows size 7×7 as cost function, since it is quite robust to a wide range of illumination changes. Additionally, we locally aggregate the costs using Adaptive support-weights [17] with radius 7 to reduce the effect of foreground fattening, but keeping the radius quite small so as not to put too much fronto-parallel assumption into the cost window. For regularization we are using two parameter sets: $\{\lambda_d = 1.0,\ \lambda_s = 0.2\}$ for the low resolution Middlebury stereo benchmark [13] and $\{\lambda_d = 0.4,\ \lambda_s = 1.0\}$ for the KITTI stereo benchmark [6] and the aerial images. The algorithm was run on a Nvidia GTX 680 GPU to which all given runtime performances relate to.

Middlebury benchmark: The Middlebury stereo benchmark [13] provides an additional discontinuity mask which we will use for the evaluation of our edge-segment based adaptive regularization. In Table 1 and Figure 5 we show the results of our algorithm both with the adaptive edge-segment regularization switched on and without. For all scenes except the teddy data set the results improve along the discontinuity regions, whereas for the teddy dataset results are worsening on the strongly slanted plane at the very bottom of the image. We are using the same parameters and cost functions described in Section 5 for all data sets and only take the gray value images of the stereo pairs as input.

KITTI benchmark: In contrast to the 4 test images of the Middlebury benchmark above, where the disparity search range is very small, the environment highly textured and the illumination conditions nearly constant, the KITTI stereo Benchmark [6] consists of 195 very challenging stereo images from ground based outdoor scenarios, together with ground truth obtained by laser scanning. In total, we achieve rank 11 in the benchmark, with a runtime of 20s per image. Additionally, we compare our results against the closest related published algorithms, also based on minimizing higher-order Total Variation (see Table 2). While we outperform the coarse-to-fine based ITGV algorithm [11] in terms of accuracy, we do not yet quite achieve the accuracy of the functional lifting based ATGV algorithm [12]. For some exemplary results of the proposed algorithm see Figure 6.

Aerial imagery: In a third data set, we apply our algorithm to aerial imagery. Despite usually having numerous overlapping images, covering every point of the scene manifold, we concentrate on showing the potential of the proposed algorithm on single stereo pairs, and apply no fusion of the resulting heightmaps in this paper. In contrast to the rectified images given in the abovementioned benchmarks, in this data set we have camera models ready for each input image, allowing us to evaluate the cost function at constant intervals in object space (using a plane-sweep approach) instead of sampling at constant disparity intervals. Thus our algorithm can treat every height value equally, whereas in disparity space, small changes in low disparities result in bigger height-differences than changes in large disparities. In Figure 7, the resulting 3D reconstruction is shown together with the two stereo images. The proposed method clearly preserves very fine details of the 3D scene (e.g. roof structures), while at the same time smoothing locally planar surfaces (church roof) quite well.

Algorithm	Tsukuba			Venus			Teddy			Cones		
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
TGV	3.66	4.33	12.0	0.21	1.00	2.88	3.93	9.66	12.1	2.44	11.1	7.20
TGV + edge	3.58	4.21	11.6	0.19	1.01	2.61	4.30	9.95	13.0	2.41	11.2	7.01

Table 1. Results of the proposed algorithm for the Middlebury Stereo benchmark (bad pixel ratio for errors $>$ 1px), once without an anisotropic diffusion tensor (TGV), once with the combined diffusion tensor of Section 2 (TGV+edge).

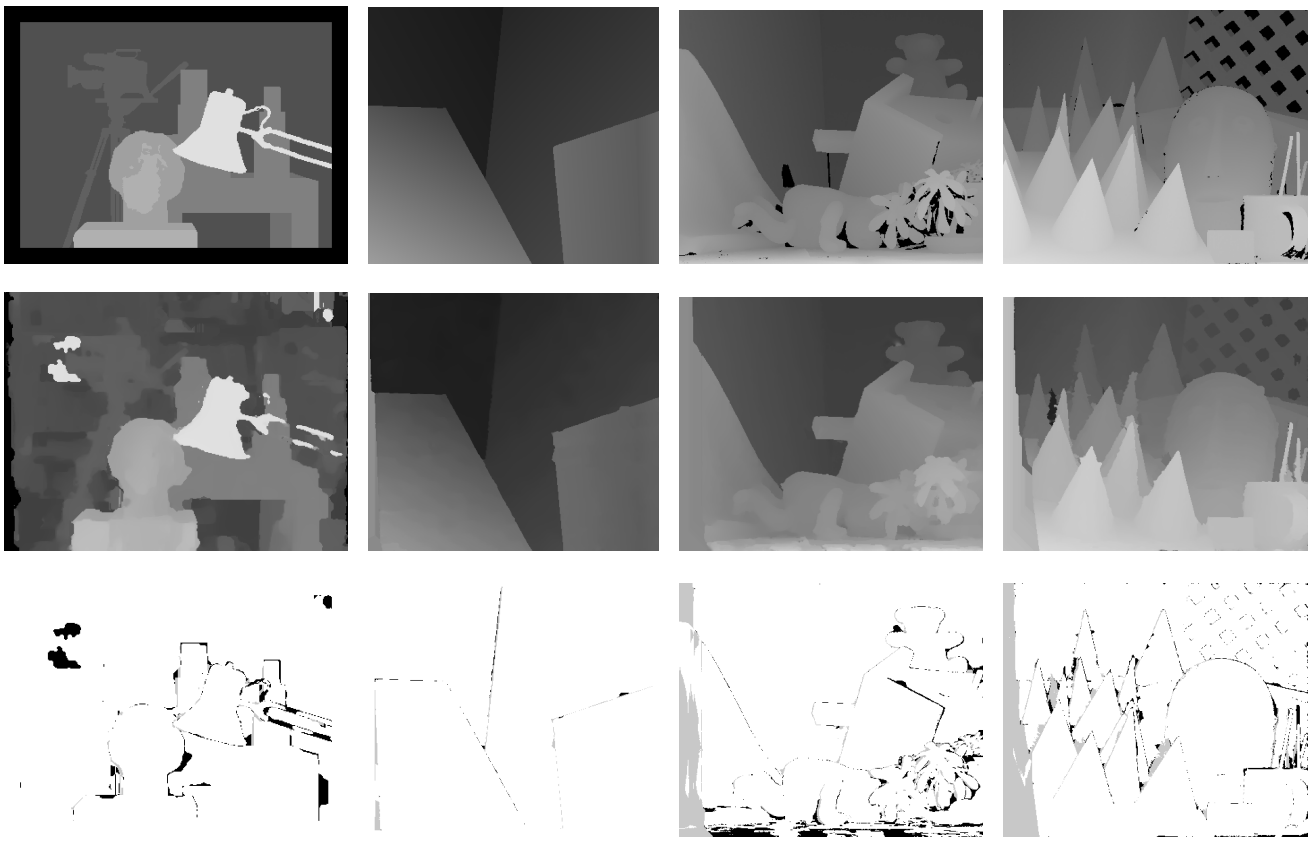
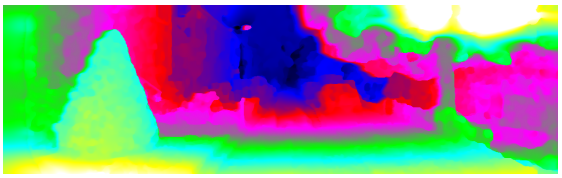


Figure 5. Results of the proposed algorithm for the Middlebury Stereo benchmark. Top row: ground truth, Middle row: our results, bottom row: bad pixel areas in black (threshold = 1px). The parameters are identical for all data sets and only the gray value images were taken.

Rank	**Method**	**Out-Noc**	**Out-All**	**Avg-Noc**	**Avg-All**	**Runtime**
8	ATGV	**5.05%**	6.91%	**1.0 px**	1.6 px	6 min
11	Proposed	5.48%	**6.60%**	1.1px	**1.2px**	20s
17	ITGV	6.31%	7.40%	1.3px	1.5px	**7s**

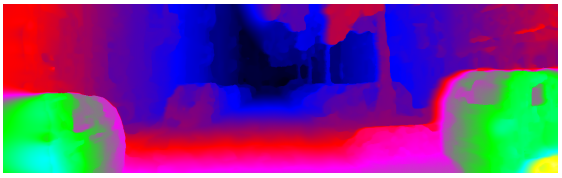
Table 2. Results for the challenging KITTI stereo benchmark [6] (195 outdoor stereo pairs). The bad pixel ratio of *Out-Noc, Out-All* is the common 3px threshold. For comparison, we further added the closest related algorithms as well. For some exemplary results see Figure 6.

(a) Reference image 1



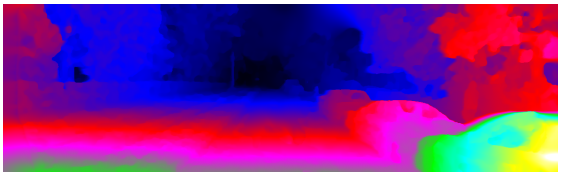
(b) Disparity map 1

(c) Reference image 7



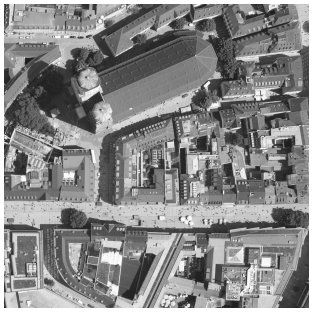
(d) Disparity map 7

(e) Reference image 15



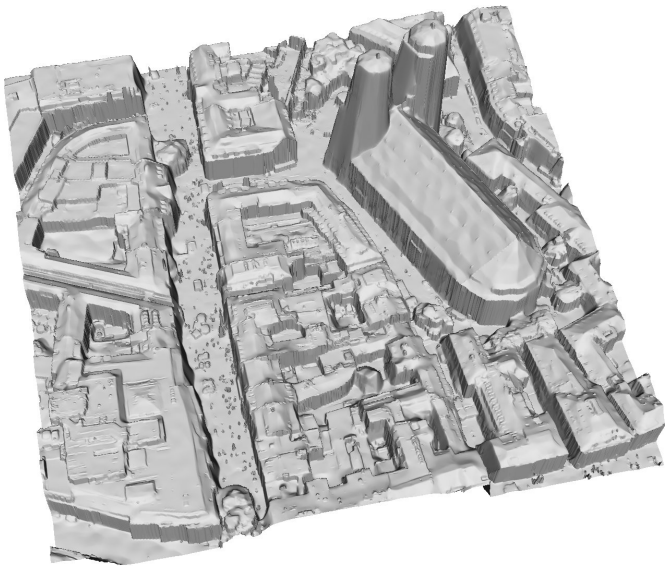
(f) Disparity map 15

Figure 6. Example results for the KITTI stereo benchmark. From top to bottom: bad, medium and good results



(a) Left image

(b) Right image



(c) Stereo reconstruction

Figure 7. a), b) Two wide-baseline aerial images (≈15cm ground resolution) c) Resulting heightmap (in camera coordinate system, not in orthogonal UTM coordinate system) of two-view stereo estimation using the proposed algorithm. Please note the fine roof structures in the 3D reconstruction, but the outliers due to moving people as well. The computation time for a 1000×1000 image using 100 disparity values is about 10s (using a Nvidia GTX 680 GPU).

6. Conclusion

In this paper we proposed an algorithm for large-scale high-accuracy stereo reconstruction of man-made worlds. To this end, we combine a non-convex data term which is robust to real-world illumination changes with a regularizer which exploits the fact that man-made worlds (buildings, cities, etc.) exhibit a large number of planar facades. The regularizer is an adaptive second-order total generalized variation modulated by means of an edge-indicator. We propose an optimization scheme consisting of a quadratic decoupling combined with an augmented Lagrangian approach which alternatingly solves the problems of correspondence finding and structure-adaptive regularization. Experiments show that the proposed augmented Lagrangian approach is faster by about a factor of 2. Validations on established stereo benchmarks and large-scale aerial images show that the proposed method provides substantial improvements over the standard TGV regularization leading to highly-accurate reconstruction of large-scale scenes.

References

[1] D. P. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods. *Computer Science and Applied Mathematics, Boston: Academic Press, 1982*, 1, 1982. 2, 4

[2] K. Bredies, K. Kunisch, and T. Pock. Total Generalized Variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010. 1

[3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High Accuracy Optical Flow Estimation Based on a Theory for Warping. *Computer Vision-ECCV 2004*, pages 25–36, 2004. 2

[4] J. Canny. A Computational Approach to Edge Detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):679–698, 1986. 2

[5] A. Chambolle and T. Pock. A First-order Primal-dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, pages 1–26, 2011. 4

[6] A. Geiger, P. Lenz, and R. Urtasun. Are we Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012. 5, 6

[7] H.-H. Nagel and W. Enkelmann. An Investigation of Smoothness Constraints for the Estimation of Displacement Vector Fields from Image Sequences. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (5):565–593, 1986. 2

[8] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense Tracking and Mapping in Real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327. IEEE, 2011. 2

[9] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers. A Convex Formulation of Continuous Multi-label Problems. *Computer Vision - ECCV 2008*, pages 792–805, 2008. 1, 2

[10] T. Pock, L. Zebedin, and H. Bischof. TGV-Fusion. *Rainbow of computer science*, pages 245–258, 2011. 1

[11] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof. Pushing the Limits of Stereo using Variational Stereo Estimation. In *Intelligent Vehicles Symposium (IV), 2012 IEEE*, pages 401–407. IEEE, 2012. 1, 2, 5

[12] R. Ranftl, T. Pock, and H. Bischof. Minimizing TGV-Based Variational Models with Non-convex Data Terms. In *Scale Space and Variational Methods in Computer Vision*, pages 282–293. Springer, 2013. 5

[13] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense two-frame Stereo Correspondence Algorithms. *International journal of computer vision*, 47(1):7–42, 2002. 5

[14] F. Steinbruecker, T. Pock, and D. Cremers. Large Displacement Optical Flow Computation without Warping. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1609–1614. IEEE, 2009. 1, 2

[15] R. G. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. LSD: A Fast Line Segment Detector with a False Detection Control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):722–732, 2010. 2

[16] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 Optical Flow. In *Proceedings of the British machine vision conference*, volume 34, pages 1–11. Citeseer, 2009. 1, 2

[17] K. Yoon and I. Kweon. Adaptive Support-weight Approach for Correspondence Search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):650–656, 2006. 5

[18] R. Zabih and J. Woodfill. Non-parametric Local Transforms for Computing Visual Correspondence. *Computer Vision - ECCV 1994*, pages 151–158, 1994. census transform, rank transform. 5