CP-Census: A Novel Model for Dense Variational Scene Flow from *RGB-D* Data

David Ferstl
ferstl@icg.tugraz.at
Gernot Riegler
riegler@icg.tugraz.at
Matthias Ruether
ruether@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

Institute for Computer Graphics and Vision,
Graz University of Technology,
Graz, Austria

Abstract

We present a novel method for dense variational *scene flow* estimation based a multi-scale *Ternary Census Transform* in combination with a patchwise *Closest Points* depth data term. On the one hand, the *Ternary Census Transform* in the intensity data term is capable of handling illumination changes, low texture and noise. On the other hand, the patchwise *Closest Points* search in the depth data term increases the robustness in low structured regions. Further, we utilize higher order regularization which is weighted and directed according to the input data by an anisotropic diffusion tensor. This allows to calculate a dense and accurate flow field which supports smooth as well as non-rigid movements while preserving flow boundaries. The numerical algorithm is solved based on a primal-dual formulation and is efficiently parallelized to run at high frame rates. In an extensive qualitative and quantitative evaluation we show that this novel method for *scene flow* calculation outperforms existing approaches. The method is applicable to any sensor delivering dense depth and intensity data such as Microsoft Kinect or Intel Gesture Camera.

1 Introduction

The structure and 3D motion of objects are essential to characterize and understand a dynamic scene. While structure from motion (*SfM*) on static scenes is well understood, non-rigid scenes still pose a challenging problem, commonly addressed as Scene Flow (*SF*). The applications for *SF* analysis range from driver assistance, surveillance, action recognition, tracking, segmentation, 3D reconstruction to camera pose estimation.

In the last decades a lot of work has addressed pure two-dimensional flow, namely Optical Flow (*OF*) [10]. The estimation of 3D motion is a relatively new topic of research. A popular way to estimate *SF* through *OF* is to use a calibrated and synchronized multi-view setup, where the *OF* estimation is combined with a depth reconstruction, as shown in [1, 13, 24, 25, 26]. With recent range sensor developments, direct depth measurements are a popular alternative to multi-view depth imaging. Such novel sensors *e.g.* Microsoft Kinect or

Intel Gesture Camera already reached a sufficient level of accuracy and robustness to allow a wide usage in the mass market. With the help of these very affordable sensors it is no longer necessary to reconstruct the whole scene through a computationally expensive multi-view setup but directly access dense depth data from the sensor.

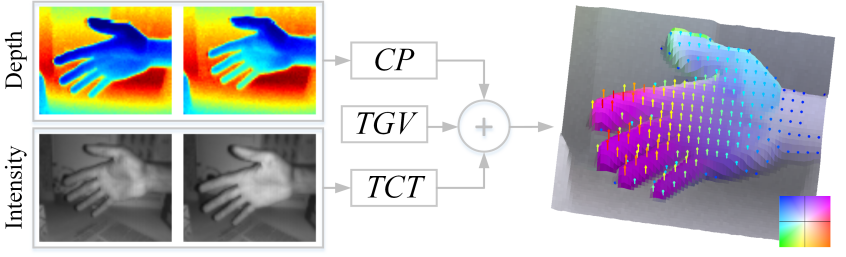


Figure 1: The *scene flow* is estimated from two consecutive depth and intensity acquisitions. The depth data term is calculated as patchwise Closest Point (*CP*) search and the intensity data term is calculated as Ternary Census Transform (*TCT*). For regularization we propose an anisotropic Total Generalized Variation (*TGV*). The flow is visualized as a color coded X,Y map (motion key in the bottom right). The Z component is shown as arrows colored according to their magnitude.

In our approach we use the output of these sensors to calculate dense *SF*, utilizing a combination of depth and intensity data as shown in Fig. 1. Our idea is to establish novel *SF* constraints to model motion through projections and image warping *directly* in 3D. In particular, we propose an intensity data term to estimate the scene correspondences by a Ternary Census Transform (*TCT*) on a local neighborhood and a depth data term to match the depth measurements directly in 3D by a patchwise Closest Point (*CP*) search. Compared to traditional pointwise constancy terms our method is invariant to most illumination changes, more robust to acquisition noise and delivers better guidance in regions with low structure or low texture. The *SF* constraints are combined with a higher order regularization term, namely Total Generalized Variation (*TGV*). Further, the regularizer is weighted and directed by an anisotropic diffusion tensor based on the input data.

The main contributions of this work are threefold: 1) We propose a novel *SF* model using advanced data terms using a *TCT* in the intensity and a patchwise 3D *CP* search in the depth data term. 2) In these constraints the flow is estimated by a warping through a projection and backprojection in 3D. 3) We formulate the proposed non-convex data terms combined with an anisotropic higher order regularization as variational energy minimization problem which is efficiently solved by the primal-dual algorithm.

2 Related Work

The first definition of the terminology of Scene Flow (*SF*) was given by Vedula *et al.* [23, 24] to estimate the 3D motion from a multi-view image sequence. Following this multi-view approach a lot of follow up work has been done, such as [1, 13, 25, 26].

With the recent availability of affordable depth sensors, methods for *SF* calculations from combined depth and intensity acquisitions have emerged. A local approach was introduced by Hadfield and Bowden [8, 9], where the *SF* calculation was modeled using a particle filter. They argue that this particle based estimation avoids oversmoothing in the flow field. Similar, Quiroga *et al.* [17] proposed a method to directly calculate the *SF* in a Lukas Kanade

(*LK*) framework. In [13] they embedded this model in a dense optimization framework. The estimation of a dense flow field in a linear optimization scheme was proposed by Letouzey *et al.* [14]. They combined an intensity constraint together with a sparse set of depth correspondences calculated through SIFT feature matching. Gottfried *et al.* [7] proposed a method for depth camera calibration to estimate dense Optical Flow (*OF*) together with a depth flow estimation. Similar, Zhang *et al.* [30] combined a global energy optimization and a bilateral filter to detect occlusions in a two-step framework. Using depth and intensity information Herbst *et al.* [8] showed how to generalize variational *OF* algorithms for *SF* calculation. They further show how *SF* aids object segmentation from motion. Similar to our work, Hornáček *et al.* [12] recently showed the advantages of estimating 3D motion directly through a patch matching in the point cloud. Unlike our method, they estimate this motion by a full rigid-body estimation for each patch using a *RGB-D* PatchMatch algorithm [2, 11]. This is especially useful for large motion, but is less capable of input noise or illumination changes.

Existing particle based approaches such as [8, 9] estimate *SF* on a sparse set of corresponding points, but these approaches deliver only a dense flow field after resampling. Other approaches such as [14] calculate local feature correspondences for depth images and a global flow estimation based on the intensity information separately. Hence, it will inevitably fail for wrong correspondence estimates. Our model builds on the success of global optimization methods as shown in [5, 7, 13, 30]. But unlike our method, these methods estimate the flow through pixelwise brightness and depth constancy. In contrast our model calculates the intensity fidelity by a patchwise Ternary Census Transform (*TCT*) on multiple scales, which is itself inherently invariant with respect to brightness changes. Further, the depth fidelity is directly calculated with the 3D point cloud by calculating the patchwise distance to the corresponding Closest Point (*CP*) estimates, similar to Iterative Closest Point (*ICP*), which makes it more accurate in low structured regions and more robust to acquisition noise. For regularization most current methods use first order penalization with a squared *L2* or a Charbonnier norm. In contrast, we use a higher order regularization with *L1* penalizer to avoid oversmoothing and flow-flattening. Edge preserving properties and smooth transitions like rotations or non-rigid movements are still possible. Furthermore, we use an anisotropic diffusion tensor based on the depth images that not only weights the flow gradient but also orients the gradient direction during the optimization process.

3 Method

The fundamental goal of 3D motion estimation is to calculate a metric motion of acquired 3D scene points in time. Consider the consecutive acquisition of a scene at time instances $t = \{1,2\}$ resulting in two consecutive depth and intensity image pairs D_1, I_1 and $D_2, I_2 \colon (\Omega \subseteq \mathbb{R}^2) \mapsto \mathbb{R}$. The scene points are observed through projections at image positions $\mathbf{x} = [x,y]^T \in \Omega$ with depth $D_t(\mathbf{x})$. Each scene point is therefore given by $\mathbf{X}_t = K^{-1}\mathbf{x}^h D_t(\mathbf{x})$, where K is the camera projection matrix and x^h denotes the homogeneous image position. The instantaneous motion for each scene point over time is given by $\mathbf{u} = \left[\frac{\mathrm{d}X}{\mathrm{d}t}, \frac{\mathrm{d}Y}{\mathrm{d}t}, \frac{\mathrm{d}D}{\mathrm{d}t}\right]^T = [u_X, u_Y, u_D]^T$. Hence, the scene motion is calculated as

$$\mathbf{X}_2 = \mathbf{X}_1 + \mathbf{u}, \quad \Longleftrightarrow \quad \begin{bmatrix} X \\ Y \\ D \end{bmatrix}_2 = \begin{bmatrix} X \\ Y \\ D \end{bmatrix}_1 + \begin{bmatrix} u_X \\ u_Y \\ u_D \end{bmatrix}, \tag{1}$$

as shown in Fig. 2. To get a correspondence in the image space the 3D movement is back projected in the image space by

$$\mathbf{x}_2 = W(\mathbf{x}_1, \mathbf{u}) = \frac{K(K^{-1}\mathbf{x}_1^h D_1(\mathbf{x}_1) + \mathbf{u})}{D_1(\mathbf{x}_1) + u_D}. \tag{2}$$

In the following we describe our variational model to estimate the general geometric flow between two frames in 3.1 and the algorithmic details to solve this minimization objective in 3.2.

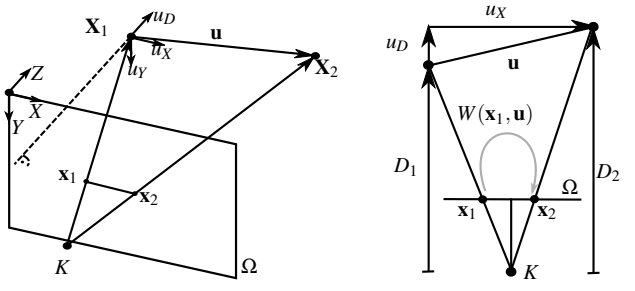


Figure 2: Flow Geometry. A scene point $\mathbf{X}_1$ acquired in the first frame moves to $\mathbf{X}_2$ in the second frame, as shown in (a). This 3D movement between two acquisitions is defined as flow $\mathbf{u}$. The projection in the image space from point $\mathbf{x}_1$ to $\mathbf{x}_2$ is defined as the warping $W(\mathbf{x}_1, \mathbf{u})$. A projection of (a) in Y direction is shown in (b).

3.1 Variational Optimization Model

The Scene Flow (*SF*) estimation in our approach is formulated as a general variational problem

$$\min_{\mathbf{u}} G_I(I_1, I_2, \mathbf{u}) + G_D(D_1, D_2, \mathbf{u}) + R(\mathbf{u}), \tag{3}$$

where the functions $G_I(I_1, I_2, \mathbf{u})$ and $G_D(I_1, I_2, \mathbf{u})$ are measuring the intensity and the depth data fidelity. Since the *SF* estimation is ill-posed we add constraints on noise and constancy expressed as a regularization force $R(\mathbf{u})$.

Traditional intensity data terms are calculated as pixelwise temporal derivatives by minimizing $I_{\Delta t}(\mathbf{x}_1, \mathbf{u}) = I_2(W(\mathbf{x}_1, \mathbf{u})) - I_1(\mathbf{x}_1)$. In contrast, we use a more complex data term which is invariant with respect to illumination and more robust to noise, namely the Ternary Census Transform (*TCT*) [21, 29]. The intensity fidelity is measured by first computing the *TCT* signature of the intensity images followed by subsequent comparison using the pixelwise Hamming distance between the images. The signature is given by

$$C(I_t; \mathbf{x}) = \bigotimes_{\mathbf{y} \in \mathcal{N}(x), \mathbf{x} \neq \mathbf{y}} \{\xi(I_t, \mathbf{x}, \mathbf{y})\}, \qquad \forall \mathbf{x} \in \Omega. \tag{4}$$

Let $\bigotimes$ denote a concatenation and $\mathcal{N}(x)$ some local neighborhood around $\mathbf{x}$. The pixelwise sign ξ is denoted by

$$\xi(I, \mathbf{x}, \mathbf{y}) = \begin{cases} 0, & \text{if } I(\mathbf{y}) - I(\mathbf{x}) + \varepsilon < 0 \\ 1, & \text{if } |I(\mathbf{y}) - I(\mathbf{x})| - \varepsilon \leq 0 \\ 2, & \text{if } I(\mathbf{y}) - I(\mathbf{x}) - \varepsilon > 0. \end{cases} \tag{5}$$

These results in a Census transformed image with a ternary string of length $|\mathcal{N}|-1$ for each pixel to encode the illumination invariant local structure. The similarity between two strings is given by the Hamming distance. In our method the intensity data term for one pixel is given as the normalized sum of differences between the census strings of the warped image $I_2(W(\mathbf{x}_1,\mathbf{u}))$ and I_1:

$$G_I(\mathbf{x},\mathbf{u}) = \frac{1}{|\mathcal{N}|-1}\sum_{i=1}^{|\mathcal{N}|-1} 1-[C_i(I_2,W(\mathbf{x},\mathbf{u})) = C_i(I_1,\mathbf{x})], \qquad \forall \mathbf{x}\in\Omega. \tag{6}$$

The Hamming distance between the Ternary Census transformed images has the nice property that it is invariant to most changes of acquisition noise and also to changes of the global illumination.

The depth fidelity is calculated directly by matching the 3D points. Traditional depth data terms, as shown in [18], are calculated by minimizing the pixelwise temporal derivatives in the depth image space $D_{\Delta t}(\mathbf{x}_1,\mathbf{u}) = D_2(W(\mathbf{x}_1,\mathbf{u})) - D_1(\mathbf{x}_1) - u_D$. These traditional pixelwise constraints fail in homogeneous regions with low depth structure. We propose a flow error metric based on the Iterative Closest Point (*ICP*) algorithm [31] which is calculated as patchwise point differences directly in 3D space to match the local surface structure. Since we have two depth acquisitions with known camera intrinsics we can calculate the residual error between both point clouds $\mathbf{X}_1$ and $\mathbf{X}_2$ according to the flow $\mathbf{u}$ by

$$G_D(\mathbf{x},\mathbf{u}) = \frac{1}{|\mathcal{N}|}\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} \|\mathbf{X}_2(\mathbf{y}) - \mathbf{u}(\mathbf{y}) - \mathbf{X}_1(\mathbf{y}^*)\|_2, \qquad \forall \mathbf{x}\in\Omega, \tag{7}$$

where the point $\mathbf{X}_1(\mathbf{y}^*)$ is denoted as the optimal correspondence to $\mathbf{X}_2(\mathbf{y})$, which in the context of *SF* is the closest point to the transformed $\mathbf{X}_2(\mathbf{y})$, *i.e.*

$$\mathbf{y}^* = \underset{\mathbf{y}\in\Omega}{\operatorname{argmin}} \|\mathbf{X}_2(\mathbf{x}) - \mathbf{u}(\mathbf{x}) - \mathbf{X}_1(\mathbf{y})\|_2. \tag{8}$$

When matching 3D patches instead of depth pixel values the method gets more robust against homogeneous depth regions and acquisition noise. Further, the direct matching in 3D does not lead to information loss due to back-projection and interpolation into the image space.

Both the intensity (6) as well as the depth fidelity term (7) are highly non-convex in the argument $\mathbf{u}$. Hence, a simple linearization of the data term is not longer sufficient. We therefore propose to perform a direct second-order Tailor expansion of the pointwise data terms around an initial flow field $\mathbf{u}_0$, similar to [28]. This is defined by

$$G_i(\mathbf{x},\mathbf{u}) \approx \tilde{G}_i(\mathbf{x},\mathbf{u}) = G_i(\mathbf{x},\mathbf{u}_0) + \nabla G_i(\mathbf{x},\mathbf{u}_0)^T(\mathbf{u}-\mathbf{u}_0) + (\mathbf{u}-\mathbf{u}_0)^T\nabla^2 G_i(\mathbf{x},\mathbf{u}_0)(\mathbf{u}-\mathbf{u}_0), \tag{9}$$

where $i=\{D,I\}$. The second derivative of the data term $\nabla^2 G_i(\mathbf{x},\mathbf{u}_0)$ must be, by definition, a positive semi-definite matrix to ensure convexity. For simplicity we use a positive semi-definite approximation of the Hessian matrix, defined as

$$\nabla^2 G(\mathbf{x},\mathbf{u}_0) = \begin{bmatrix} G_{xx}(\mathbf{x},\mathbf{u}_0)^+ & 0 & 0 \\ 0 & G_{yy}(\mathbf{x},\mathbf{u}_0)^+ & 0 \\ 0 & 0 & G_{zz}(\mathbf{x},\mathbf{u}_0)^+ \end{bmatrix}, \tag{10}$$

where only positive second order derivatives are allowed while mixed derivatives are neglected. In [27] it has been shown that this approximation will not harm the estimation accuracy.

For regularization most of the common Optical Flow (*OF*) and *SF* approaches use a first order regularizer or non-local variation with *L1* or *L2* penalizer. While the quadratic (*L2*) penalization leads to over-smoothed results, the Total Variation (*TV*) semi-norm (*L1*) enforces a piecewise constancy in the flow field, which is only valid for simple translational movements. To avoid these disadvantages we use a higher order model, namely the Total Generalized Variation (*TGV*) of second order, proposed by Bredies *et al.* [3]. It not only includes the first derivative but also the second order derivatives to approximate the *SF* by piecewise affine transformations. The primal definition of the second order *TGV* is formulated as

$$R(\mathbf{u},\mathbf{v}) = \min_{\mathbf{u},\mathbf{v}} \left\{ \alpha_1 \int_\Omega |\nabla\mathbf{u} - \mathbf{v}| \,\mathrm{d}x + \alpha_0 \int_\Omega |\nabla\mathbf{v}| \,\mathrm{d}x \right\}, \tag{11}$$

where $\mathbf{u}$ denotes the flow field and $\mathbf{v}$ its first derivative. The scalars $\alpha_0, \alpha_1 \in \mathbb{R}$ are used to weight each order.

Although the *TGV* shows edge preserving capabilities we additionally use the edge information from our input depth images to refine this regularization. Henceforth, we include an anisotropic diffusion tensor $T^{\frac{1}{2}}$, known as the Nagel-Enkelmann operator [15], based on the input depth image D_1. This tensor is calculated by $T^{\frac{1}{2}} = \exp\left(-\beta \left|\nabla D_1\right|^\gamma\right) \mathbf{n}\mathbf{n}^\mathrm{T} + \mathbf{n}^\perp\mathbf{n}^{\perp\mathrm{T}}$, where $\mathbf{n} = \frac{\nabla D_1}{|\nabla D_1|}$ is the normalized direction of the depth image gradient and $\mathbf{n}^\perp$ is the normal vector to the gradient. The scalars $\beta, \gamma \in \mathbb{R}$ adjust the magnitude and the sharpness of the tensor. The anisotropic diffusion tensor not only weights the motion gradient but also orients the gradient direction during the optimization process. This regularization term has shown great success in stereo reconstruction [19] and depth image upsampling [6].

With the combination of convex *TGV* regularization and anisotropic weighting we achieve smooth transitions between flows, typically occuring at object rotations and non-rigid movements, while sharp flow boundaries between moving objects can still be preserved.

Based on our definitions, the final energy in our optimization model is defined as

$$\min_{\mathbf{u},\mathbf{v}} \left\{ \lambda_I \int_\Omega w|\tilde{G}_I| \,\mathrm{d}x + \lambda_D \int_\Omega w|\tilde{G}_D| \,\mathrm{d}x + \alpha_1 \int_\Omega |T^{\frac{1}{2}}(\nabla\mathbf{u} - \mathbf{v})| \,\mathrm{d}x + \alpha_0 \int_\Omega |\nabla\mathbf{v}| \,\mathrm{d}x \right\}, \tag{12}$$

where $\lambda_D, \lambda_I \in \mathbb{R}$ are parameters to weight the individual data terms. The pixelwise confidence for the data measurements is given by $w\colon \Omega \mapsto [0,1]$. This confidence is set to 0 where no depth measurements are available, *e.g.* for stereo sensors at occluded regions, otherwise it is 1.

3.2 Numerical Implementation

The proposed optimization problem is convex but non-smooth due to the *L1* norms and the possibility of zeros in the weighting parameter w. Therefore, the optimization of this problem is not a trivial task. Since (12) is convex in $\mathbf{u}$ and $\mathbf{v}$ we can make use of the dual principle. After introducing Lagrange multipliers for the constraints and biconjugation using the Legendre Fenchel transform (*LF*) we are able to reformulate the problem as a convex-concave saddle point problem discretized on a Cartesian image grid of size $M \times N$.

This saddle point problem is solved using the primal-dual optimization scheme proposed in [4]. This scheme provides a fast convergence rate and is parallelized in the implementation resulting in high frame rates. The local approximation of the *SF* constraints are only valid for small displacements. Therefore, the primal-dual calculation is embedded into a coarse-to-fine framework. We employ image pyramids with a downsampling factor of $\nu = 0.8$ for this purpose. The iterative optimization runs in 5 warps with 50 iterations per level. Due to warping in 3D space the camera projection matrix has to be adjusted according to each pyramid level by $K_i = K\text{diag}([\nu^i, \nu^i, 1]), \forall i = 0...L$, where K_i is the camera matrix at level i and $i = 0$ is the finest level. The solution of each level is propagated to the next finer level starting with $\mathbf{u}_L = \mathbf{0}$, $\mathbf{v}_L = \mathbf{0}$ at the coarsest level. The weighting parameters for all terms in our energy are kept constant over all levels. The first and second order derivatives are computed using standard central differences. Instead of an exhaustive search we use a standard *kd-tree* for patchwise Closest Point (*CP*) search to increase the computation speed, where the patch size is set to 5×5. The *TCT* term is calculated using the minimum over multiple window sizes ranging from 5×5 to 11×11. Due to lack of space, we will outline the detailed numerical optimization scheme in the supplementary material.

4 Evaluation

In this section we provide an extensive qualitative and quantitative evaluation of our method, which we further address as *CP-Census*. An analysis of the properties and the effects of different terms in our method on real and synthetic datasets is shown in 4.1. A numerical evaluation in terms of speed and accuracy compared to state of the art (*SOTA*) Scene Flow (*SF*) methods is given in 4.2. For visual real world evaluations we used a PMD Nano Time of Flight (*ToF*) camera [16] and a Microsoft Kinect for Windows v2 camera (K4Wv2)[a]. Following [9] and [12], the flow error is calculated with the commonly used error measurements Average Angular Error (*AAE*) and the End Point Error (*EPE*) in 2D and 3D space. For the Middlebury evaluation we additionally provide the disparity error RMS_{V_Z}. The average runtime over all experiments is 1.47s computed on a Nvidia GTX680 GPU. Further visual evaluations are shown in the supplemental material.

4.1 Scene Flow Evaluation on Synthetic and Real Datasets

In this experiment we quantitatively evaluate our *SF* algorithm and the contributions of the individual terms in our objective function compared to *SOTA* Optical Flow (*OF*) and *SF* algorithms. We use a synthetic and a real dataset consisting of moving objects in a static scene. In the synthetic dataset a cube is rotated and translated in front of a static background. The scene was generated including depth and intensity image pairs as well as a groundtruth *SF*. To simulate acquisition noise we applied Gaussian noise to the input data. To compare the effects of different object movements we show the errors for a pure translation of 20% of the object size in X direction (T_X), a pure translation of 20% towards the camera (T_Z) and a rotation of 15 degrees about the Z axis (R_Z).

In Table 1 we compare our results with two *SOTA OF* methods, namely *NL-TV-NCC* methods of Werlberger *et al.* [28] and *Classic-NL-Full* methods of Sun *et al.* [22] and the currently best performing *SF* method of Hornáček *et al.* [12].

[a]The K4Wv2 developer kit is preliminary software and/or hardware and APIs are preliminary and subject to change.

	$T_X = 20\%$		$T_Z = -20\%$		$R_Z = 15°$	
	EPE_{SF}	AAE_{SF}	EPE_{SF}	AAE_{SF}	EPE_{SF}	AAE_{SF}
NL-TV-NCC	0.282	5.61	0.191	3.07	0.291	5.06
Classic+NL-Full	0.260	4.57	0.303	3.29	0.388	5.68
Hornáček et al. [12]	0.089	3.85	0.090	3.30	0.056	2.43
CP-Census w/o tensor	0.096	3.92	0.117	3.32	0.088	3.25
CP-Census TVL1	0.037	1.43	**0.041**	**0.95**	0.066	1.92
CP-Census	**0.035**	**1.42**	0.042	**0.95**	**0.053**	**1.67**

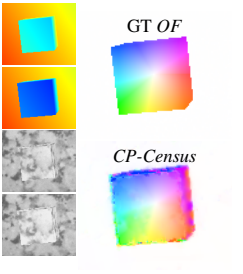


Table 1: Scene Flow evaluation on a synthetic dataset. Comparison of our method with *SOTA OF* and *SF* methods at different object movements in terms of *EPE* and *AAE* in 3D. Further, evaluation results of our methods are shown, where different terms are turned off. The best result for each movement is highlighted and the second best is underlined. On the right side, the input and results of our method for pure Z movement are shown.

The 3D error of the *OF* methods is calculated by a projection into 3D space with the known depth maps. We further show a visual evaluation on real acquisitions of freely moving rigid and non-rigid objects acquired by the PMD Nano and the K4Wv2 camera in Fig. 3.

Due to the additional depth information in our model it is obvious that we significantly outperform traditional *OF* approaches. For object rotations or non-rigid movements one can see the advantage of the higher order regularization in our model. The modeling of smooth transitions is hard for first order regularization (*TVL1*) while our higher order model can cope with these types of transitions. While the first order approaches work well for pure translational movements they are not suitable to model smooth flow transitions such as rotations or non-rigid movements since it enforces piece wise constant solutions in the flow field. Furthermore the anisotropic tensor has a big impact on the quality since it bounds the flow field at object boundaries. The *RGB-D* PatchMatch approach of Hornáček *et al.* [12] delivers comparable results for the flow magnitude (*EPE*) but lacks in angular precision (*AAE*). Further it has problems at larger noise levels or illumination changes which appear at the PMD Nano sequences (Fig. 3).

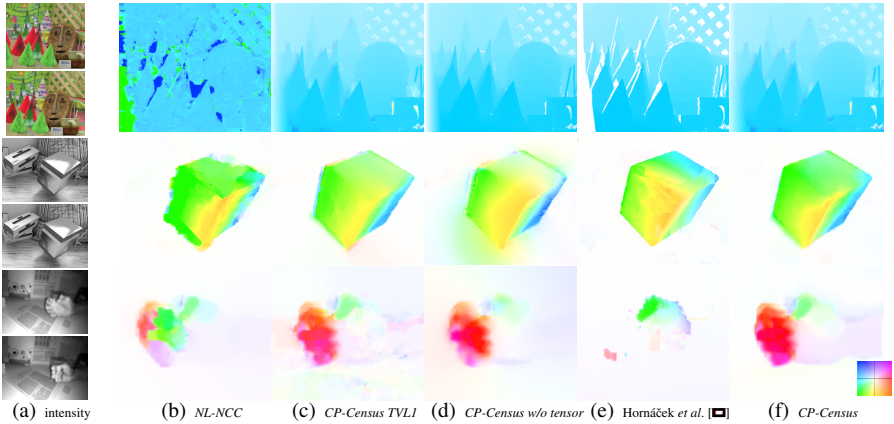
4.2 Middlebury Evaluation

In order to perform a quantitative comparison to more *SOTA SF* methods, we evaluate our method using an existing scene flow benchmark dataset. We follow [1, 8, 12, 13, 17, 18, 30], which use the rectified stereo intensity and disparity maps from the Middlebury *Cones*, *Teddy* and *Venus* datasets [20] to simulate scene flow. In this setting, two images are acquired with a pure horizontal camera movement. This allows to recover a ground truth scene motion at every point in a cluttered scene with pure X movement, where the 3D movement in Y and Z direction is zero and the movement in X-direction is given by the baseline. As in the compared methods, the disparity maps are used to simulate the output of the depth sensor. The calculated *SF* is backprojected into the image space for a direct comparison with the ground truth disparity maps. In Table 2 the evaluation results compared to several *SOTA* methods for *SF* from stereo and *SF* from depth and intensity data are shown.

What can be clearly seen is that our method delivers a *SF* quality which is superior compared to other *SOTA* methods in most cases. We deliberately use the same parameters for all three datasets even though the Venus dataset has other lighting and surface conditions.

	Cones			*Teddy*			*Venus*		
	EPE_{OF}	RMS_{Vz}	AAE_{OF}	EPE_{OF}	RMS_{Vz}	AAE_{OF}	EPE_{OF}	RMS_{Vz}	AAE_{OF}
Basha et al. [9](2 views) (st)	0.58	N/A	0.39	0.57	N/A	1.01	0.16	N/A	1.58
Huguet and Devernay [13] (st)	1.10	N/A	0.69	1.25	N/A	0.51	0.31	N/A	0.98
Hadfield and Bowden [9]	1.24	0.06	1.01	0.83	0.03	0.83	0.36	0.02	1.03
Quiroga et al. [13]	0.57	0.05	0.52	0.69	0.04	0.71	0.31	**0.00**	1.26
Hornáček et al. [14]	0.54	**0.02**	0.52	0.35	**0.01**	0.16	0.26	0.02	0.64
CP-Census	**0.40**	0.03	**0.04**	**0.31**	0.02	**0.05**	**0.15**	**0.00**	**0.41**

Table 2: Evaluation on the Middlebury dataset. The error is measured by *EPE* / *AAE* in 2D, and *RMS* in depth. The best result for each dataset is highlighted and the second best is underlined. Methods that calculate SF from stereo are marked with (st).



(a) intensity (b) *NL-NCC* (c) *CP-Census TVL1* (d) *CP-Census w/o tensor* (e) Hornáček *et al.* [14] (f) *CP-Census*

Figure 3: Evaluation of different *SF* methods on real image sequences. In (a) the input intensity images are shown. In the first row the results of the Middlebury *Cones* sequence, in the second row the flow of a rotated box with the K4Wv2 and in the third row a hand closing sequence (non-rigid movement) acquired with the PMD Nano are shown. Each scene is evaluated for *NL-NCC OF* (b) *CP-Census* without a second order regularization (c), *CP-Census* without anisotropic diffusion (d), the method of Hornáček *et al.* (e) compared to our full method (f). The motion key is shown in the bottom right of (f). Figure best viewed magnified in the electronic version.

5 Conclusion

We proposed a method for the estimation of *scene flow* from depth and intensity data. The estimation is formulated as a convex energy minimization problem using sophisticated non-convex data terms together with an anisotropic higher order regularization. Our method better handles scenes with low texture or low structure and is robust to illumination changes. Further, it can cope with smooth flow transitions, which occur at rotations or non-rigid movements, while sharp boundaries of the flow field are preserved. In a quantitative and qualitative evaluation we show that our method clearly outperforms existing state of the art approaches. As a future perspective we want to make use of the high estimation quality in applications such as non-rigid structure from motion or depth image superresolution and camera pose estimation of dynamic scenes.

Acknowledgments

This work was supported by *Infineon Technologies Austria AG* and the Austrian Research Promotion Agency (FFG) under the *FIT-IT Bridge* program, project #838513 (TOFUSION).

References

[1] Tali Basha, Yael Moses, and Nahum Kiryati. Multi-view scene flow estimation: A view centered variational approach. In *CVPR*, 2010.

[2] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo - stereo matching with slanted support windows. In *BMVC*, 2011.

[3] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. *SIAM Journ. on Imaging Sciences*, 3(3):492– 526, 2010.

[4] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120 –145, 2011.

[5] Xiaofeng Ren Evan Herbst and Dieter Fox. Rgb-d flow: Dense 3-d motion estimation using color and depth. In *ICRA*, 2013.

[6] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *ICCV*, 2013.

[7] Jens-Malte Gottfried, Janis Fehr, and ChristophS. Garbe. Computing range flow from multi-modal kinect data. In *ISVC*, 2011.

[8] Simon Hadfield and Richard Bowden. Kinecting the dots: Particle based scene flow from depth sensors. In *ICCV*, 2011.

[9] Simon Hadfield and Richard Bowden. Scene particles: Unregularized particle-based scene flow estimation. *TPAMI*, 36(3):564–576, 2014.

[10] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185 –203, 1981.

[11] Michael Hornáček, Christoph Rhemann, Margrit Gelautz, and Carsten Rother. Depth super resolution by rigid body self-similarity in 3d. In *CVPR*, 2013.

[12] Michael Hornáček, Andrew Fitzgibbon, and Carsten Rother. Sphereflow: 6dof scene flow from rgb-d pairs. In *CVPR*, 2014.

[13] Frédéric Huguet and Frédéric Devernay. A variational method for scene flow estimation from stereo sequences. In *ICCV*, 2007.

[14] Antoine Letouzey, Benjamin Petit, and Edmond Boyer. Scene flow from depth and color images. In *BMVC*, 2011.

[15] Hans-Hellmut Nagel and Wilfried Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *TPAMI*, 8(5):565 –593, 1986.

[16] *Camboard Nano*. PMD Technologies. Siegen, Germany.

[17] Julian Quiroga, Frédéric Devernay, and James L. Crowley. Scene flow by tracking in intensity and depth data. In *CVPR Workshops*, 2012.

[18] Julian Quiroga, Frédéric Devernay, and James L. Crowley. Local/global scene flow estimation. In *ICIP*, 2013.

[19] Rene Ranftl, Stefan Gehrig, Thomas Pock, and Horst Bischof. Pushing the limits of stereo using variational stereo estimation. In *IEEE Intelligent Vehicles Symposium*, 2012.

[20] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *CVPR*, 2003.

[21] Fridtjof Stein. Efficient Computation of Optical Flow Using the Census Transform. In *DAGM*, 2004.

[22] Deqing Sun, Stefan Roth, and MichaelJ. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *IJCV*, pages 1–23, 2013.

[23] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *ICCV*, 1999.

[24] Sundar Vedula, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. *TPAMI*, 27(3):475–480, 2005.

[25] Christoph Vogel, Konrad Schindler, and Stefan Roth. Piecewise rigid scene flow. In *ICCV*, 2013.

[26] Andreas Wedel, Thomas Brox, Tobi Vaudrey, Clemens Rabe, Uwe Franke, and Daniel Cremers. Stereoscopic scene flow computation for 3d motion understanding. *IJCV*, 95 (1):29–51, 2011.

[27] Manuel Werlberger. *Convex Approaches for High Performance Video Processing*. PhD thesis, Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria, June 2012.

[28] Manuel Werlberger, Thomas Pock, and Horst Bischof. Motion estimation with non-local total variation regularization. In *CVPR*, 2010.

[29] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *ECCV*, 1994.

[30] Xiaowei Zhang, Dapeng Chen, Zejian Yuan, and Nanning Zheng. Dense scene flow based on depth and multi-channel bilateral filter. In *ACCV*, 2013.

[31] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *IJCV*, 13(2):119–152, 1994.