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

Dynamic scene understanding is an essential topic in computer vision. It tries to combine information from tracking, 3D reconstruction, segmentation, motion estimation to infer information about an ever changing 3D environment. While structure from motion for measuring movements in space is well understood on static scenes, the motion estimation of non-static scenes, known as Scene Flow (*SF*), still pose a challenging problem. This gets even harder if the moving objects are non-rigid. A popular way to estimate *SF* is to use a calibrated and synchronized multi-view setup and combine traditional Optical Flow (*OF*) estimation with simultaneous 3D reconstruction [1, 4]. With the recent range sensor developments, such as the Microsoft Kinect or the Intel Gesture Camera, the *SF* estimation solely from RGB-D data became a popular alternative [2, 3, 5].

In this paper we show a novel method for accurate and robust *SF* estimation of non-rigid scenes from RGB-D data. This estimation is solved in an dense variational energy minimization framework

$$\min_{\mathbf{u}} G_I(I_1, I_2, \mathbf{u}) + G_D(D_1, D_2, \mathbf{u}) + R(\mathbf{u}) \tag{1}$$

based on a multi-scale Ternary Census Transform (*TCT*) for the intensity data term G_I in combination with a depth data term G_D based on the patch-wise Closest Point (*CP*) distance, as shown in Figure 1. The motion in our estimation is modeled as *direct* projection and image warping W in 3D.

In particular, we propose an intensity data term G_I to estimate the scene correspondences given by the *TCT* on a local neighborhood $\mathcal{N}$:

$$G_I(\mathbf{x}, \mathbf{u}) = \frac{1}{|\mathcal{N}|-1} \sum_{i=1}^{|\mathcal{N}|-1} 1 - [C_i(I_2, W(\mathbf{x}, \mathbf{u})) = C_i(I_1, \mathbf{x})], \tag{2}$$

Where C is the ternary census signature of each patch. This *TCT* term calculates the intensity difference by an encoding of the illumination invariant local structure. The similarity is calculated by the Hamming distance between the signature patches. The a depth data term G_D is calculated as the patch-wise distance to the *CP* in 3D, which makes it more robust in low structured regions and in case of acquisition noise:

$$G_D(\mathbf{x}, \mathbf{u}) = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{y} \in \mathcal{N}(\mathbf{x})} \|\mathbf{X}_2(\mathbf{y}) - \mathbf{u}(\mathbf{y}) - \mathbf{X}_1(\mathbf{y}^*)\|_2. \tag{3}$$

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 R, namely Total Generalized Variation (*TGV*). The regularizer is weighted and directed by an anisotropic diffusion tensor based on the input data. Because both the intensity as well as the depth data are highly non-convex a simple linearization as in traditional methods is not longer sufficient. We therefore perform a direct second-order Taylor expansion of the pointwise data terms, similar to [6]. The proposed whole variational energy model is efficiently solved based on the primal-dual formulation and is efficiently parallelized to run at high frame rates.

In an extensive evaluation we show the different properties and contributions of the different terms in our model. The applicability of our method to different kinds of camera modalities is shown in Figure 2. Beyond that, we show that the accuracy of our method is superior compared to current *SF* approaches based on the Middlebury Benchmark, as shown in Table 1. Our method better handles scenes with low texture or low structure and is robust to illumination changes. It can cope with smooth flow transitions, which occur at rotations or non-rigid movements, while sharp boundaries of the flow field are preserved.

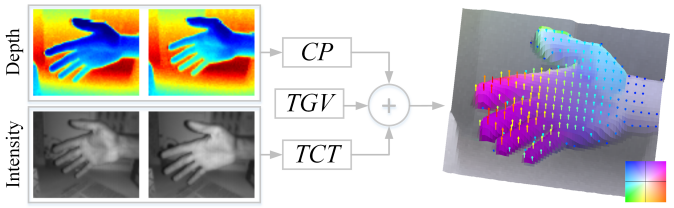
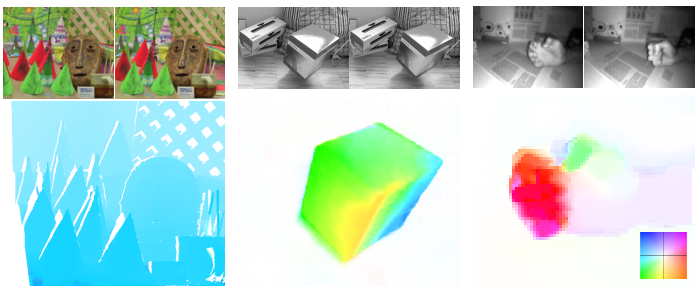


Figure 1: The *scene flow* is estimated from two consecutive depth and intensity acquisitions. The depth data term is calculated as patch-wise Closest Point (*CP*) search and the intensity data term is calculated as Ternary Census Transform (*TCT*). For regularization we propose an anisotropic Total 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.



(a) Middlebury (b) K4Wv2 (c) PMD Nano

Figure 2: *CP-Census SF* results on real image sequences. In the first column the results of the Middlebury *Cones* sequence, in the second column the flow of a rotated box with the K4Wv2 and in the third column a hand closing sequence (non-rigid movement) acquired with the PMD Nano are shown.

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

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

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

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

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

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

	Cones EPE / RMS_{V_Z} / AAE			*Teddy* EPE / RMS_{V_Z} / AAE			*Venus* EPE / RMS_{V_Z} / AAE		
Basha et al. [1](2 views) (st)	0.58	N/A	0.39	0.57	N/A	1.01	0.16	N/A	1.58
Huguet and Devernay [4] (st)	1.10	N/A	0.69	1.25	N/A	0.51	0.31	N/A	0.98
Hadfield and Bowden [2]	1.24	0.06	1.01	0.83	0.03	0.83	0.36	0.02	1.03
Quiroga et al. [5]	0.57	0.05	0.52	0.69	0.04	0.71	0.31	**0.00**	1.26
Hornáček et al. [3]	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 1: Quantitative comparison of *SF* methods on the Middlebury dataset. The error is measured by *EPE*/*AAE* in 2D, and *RMS* in Z direction. The best result for each dataset is highlighted and the second best is underlined. Methods that calculate *SF* from stereo are marked with (st).