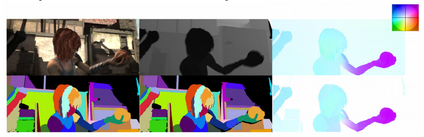
Multiframe Scene Flow with Piecewise Rigid Motion

Vladislav Golyanik, Kihwan Kim, Robert Maier, Matthias Nießner, Didier Stricker and Jan Kautz

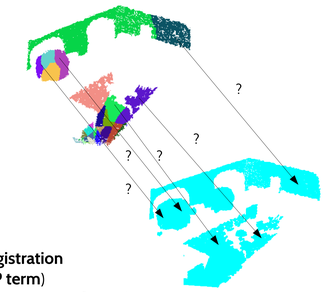
3DV 2017

Overview

input RGB-D frames (overlayed) — ground truth



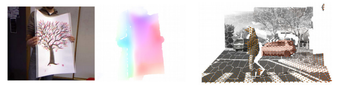
oversegmentation of the reference frame — segmentation transfer into the current frame — our MSF result (projected)



Contributions

- transformations are parametrized by **piecewise rigid** motions
- **depth channel** is used to obtain oversegmentation of the scene
- segmentation of a scene is kept **fixed**
- a global scene-flow formulation over **multiple frames**
- take advantage of **point set registration** (**projective point-to-plane ICP term**)
- **lifting function** for coherent segment transformations
- using the efficient framework of non-linear least squares

Related work



Quiroga *etal. ECCV, 2014.* — Vogel *etal. ICCV, 2013.*

Jaimez *et al. ICRA, 2015.* — Jaimez *et al. 3DV, 2015.*

Proposed Energy Functional

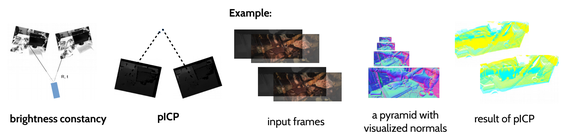
$$\mathfrak{E}(\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^{|\mathrm{Z}|}, \mathbf{w}) = \sum_{\zeta \in \mathrm{Z}} \alpha_\zeta \, \mathfrak{E}_{\mathrm{data}}(\mathbf{T}^\zeta) + \sum_{\zeta \in \mathrm{Z}} \beta_\zeta \, \mathfrak{E}_{\mathrm{pICP}}(\mathbf{T}^\zeta) \; + \gamma_\zeta \sum_{\zeta \in \mathrm{Z}} \mathfrak{E}_{\mathrm{l.reg.}}(\mathbf{T}^\zeta, \mathbf{w}) + \eta \; \mathfrak{E}_{\mathrm{r.opt.}}(\mathbf{w}) + \sum_{\zeta=3}^{|\mathrm{Z}|} \lambda_\zeta \, \mathfrak{E}_{\mathrm{c.}}(\mathbf{T}^\zeta)$$

data term (brightness constancy) + projective ICP term (point-to-plane) — lifted segment pose regularizer + robust weight optimizer — multiframe pose concatenation term

I

$$\mathfrak{E}_{\mathrm{data}}(\mathbf{T}) = \sum_k \sum_{\mathbf{p} \in \Omega_k} \left\| (\mathcal{I}_1(\mathbf{p}) - \mathcal{I}_2(\pi(g(\mathbf{T}_k, \pi^{-1}(\mathbf{p}, z_\mathbf{p}))))^2 \right\|_\epsilon$$

$$\mathfrak{E}_{\mathrm{pICP}}(\mathbf{T}) = \sum_k \sum_{\mathbf{p} \in \Omega_k} \left\| ((g(\mathbf{T}_k, \mathbf{p}) - \mathbf{p}^{\mathrm{corr}}) \cdot \mathbf{n}_\mathbf{p})^2 \right\|_\epsilon$$



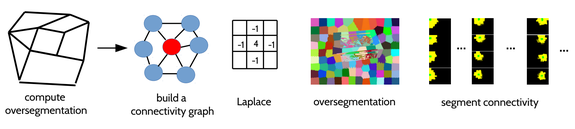
brightness constancy — pICP — Example: input frames — a pyramid with visualized normals — result of pICP

in every iteration of projective ICP:
- update per-segment correspondences
- update camera per-segment alignment

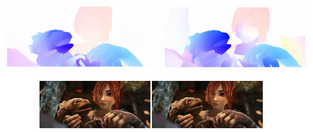
II

$$\mathfrak{E}_{\mathrm{l.\,reg}}(\mathbf{T}, \mathbf{w}) = \sum_{w_{j,h} : \Psi[j,h]=1} \left\| (w_{j,h}^2 \, (\mathbf{T}_j - \mathbf{T}_h))^2 \right\|_2$$

$$\mathfrak{E}_{\mathrm{r.\,opt.}}(\mathbf{w}) = \sum_{\forall w_{j,h} : \Psi[j,h]=1} \left\| (1 - w_{j,h}^2)^2 \right\|_2$$

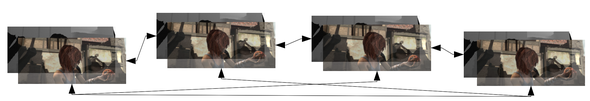


compute oversegmentation — build a connectivity graph — Laplace — oversegmentation — segment connectivity



III

$$\mathfrak{E}_{\mathrm{c.}}(\mathbf{T}^{l,m}) = \sum_k \left\| (\mathbf{T}_k - \mathbf{T}_k^{m-1,m} \cdot \ldots \cdot \mathbf{T}_k^{l,l+1})^2 \right\|_2$$



global optimization over multiple frames

Huber norm:

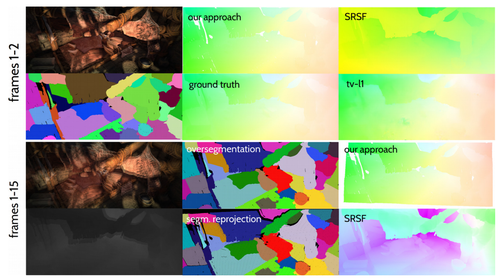
$$\left\| a^2 \right\|_\epsilon = \begin{cases} \frac{1}{2} a^2, & \text{for } |a| \leq \epsilon \\ \epsilon(|a| - \frac{1}{2}\epsilon), & \text{otherwise} \end{cases}$$

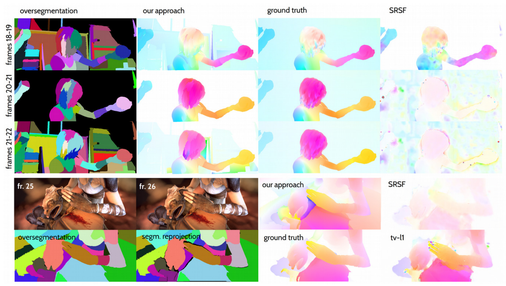
Experiments

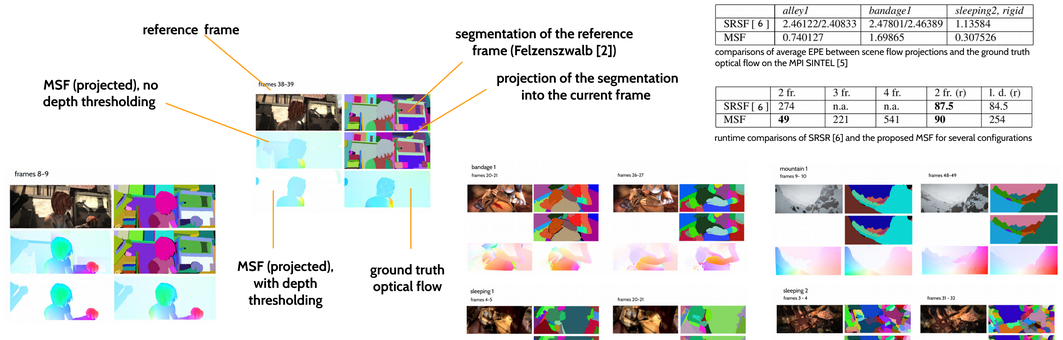
In the experimental evaluation we use:
- MPI SINTEL [5]
- virtual KITTI [3]
- Bonn multibody data set [7]
- own RGB-D recordings

... and compare the following methods:
- Primal-Dual Flow [4]
- Semi-Rigid Scene Flow [6]
- Multi-Frame Optical FLow [8]
- tv-l1 optical flow [10]

End Point Error is defined as $\|(u - u_{\mathrm{GT}}), (v - v_{\mathrm{GT}})\|$
$(u, v)^\mathsf{T}$ is a projected flow vector
$(u_{\mathrm{GT}}, v_{\mathrm{GT}})^\mathsf{T}$ is a ground truth vector





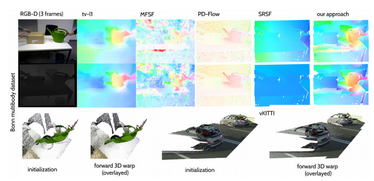


	alley1	*bandage1*	*sleeping2, rigid*
SRSF [6]	2.46122/2.40833	2.47801/2.46389	1.13584
MSF	0.740127	1.69865	0.307526

comparisons of average EPE between scene flow projections and the ground truth optical flow on the MPI SINTEL [5]

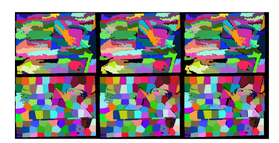
	2 fr.	3 fr.	4 fr.	2 fr. (r)	1. d. (r)
SRSF [6]	274	n.a.	n.a.	**87.5**	84.5
MSF	**49**	221	541	**90**	254

runtime comparisons of SRSR [6] and the proposed *MSF* for several configurations

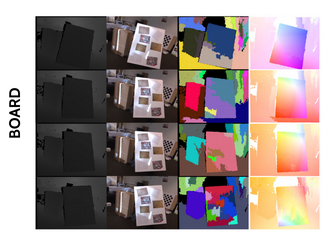


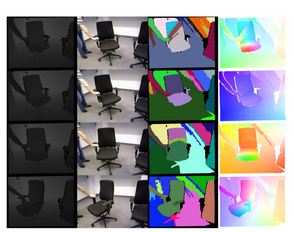
Results of several RGB-D scene flow and optical flow approaches on the Bonn multibody data set [7] and an example of an urbane driving scene processing form the vKITTI data set [3] by MSF (bottom right)

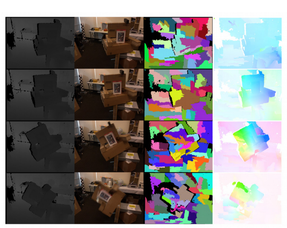
Segmentation transfer from the reference frame to three other frames in *alley1*



Segmentation transfer on the Bonn watering can sequence with Felzenszwalb segmentation [2] (top row) and SLIC segmentation [1] (bottom row)







References

[1] R. Achanta *etal.* Slic superpixels compared to state-of-the-art superpixel methods. T-PAMI, 2012.
[2] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. IJCV, 2004.
[3] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In CVPR, 2016.
[4] M. Jaimez, M. Souiai, J. González-Jiménez, and D. Cremers. A primal-dual framework for real-time dense rgb-d scene flow. In ICRA, 2015.
[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In ECCV, 2012.
[6] J. Quiroga, T. Brox, F. Devernay, and J. L. Crowley. Dense semi-rigid scene flow estimation from RGBD images. In ECCV, 2014.
[7] J. Stueckler and S. Behnke. Efficient dense rigid-body motion segmentation and estimation in rgb-d video. IJCV, 2015.
[8] B. Taetz, G. Bleser, V. Golyanik, and D. Stricker. Occlusion-aware video registration for highly non-rigid objects. In WACV, 2016.
[9] C. Vogel, K. Schindler, and S. Roth. Piecewise rigid scene flow. In ICCV, 2013.
[10] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In GCPR, 2007.

Energy Optimization

We minimize the energy using **Levenberg-Marquardt** (*ceres* solver, C++, multithreaded):

$$\mathrm{x}' = \arg\min_{\mathrm{x}} \|\mathbf{F}(\mathrm{x})\|_2^2 \qquad \mathbf{F}(\mathrm{x} + \Delta\mathrm{x}) \approx \mathbf{F}(\mathrm{x}) + \mathbf{J}(\mathrm{x})\Delta\mathrm{x} \qquad \left(\mathbf{J}(x)^\mathsf{T}\mathbf{J}(\mathrm{x}) + \lambda \, \mathrm{diag}(\mathbf{J}(x)^\mathsf{T}\mathbf{J}(\mathrm{x}))\right) \Delta\mathrm{x} = -\mathbf{J}(\mathrm{x})^\mathsf{T}\mathbf{F}(\mathrm{x})$$

The total number of residuals:

$$\mathrm{M} = {}^{\mathrm{N}}C_2 \, (n_\mathcal{C} + n_\mathcal{D}) + ({}^{\mathrm{N}}C_2 + 1) \, n_{\mathrm{pp}} + n_{\mathrm{c}}$$

$$n_{\mathrm{pp}} = {}^{\mathrm{N}}C_2 \sum \psi_{k,l}$$

$$n_{\mathrm{c}} = \mathrm{K} \, (\mathrm{N} - 2)$$

- # of pose pairs (number of non-zero elements in the segment adjacency matrix), for every two frame case
- # of concatenated transformations
- # of individual frame-to-frame (two frame) combinations
- # of valid pixel pairs (two frames)
- # of valid 3D point pairs (two frames)
- # of weights (lifting)