Estimating Vehicle Ego-Motion and Piecewise Planar Scene Structure from Optical Flow in a Continuous Framework

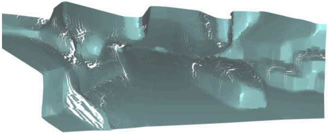
Andreas Neufeld, Johannes Berger, Florian Becker, Frank Lenzen and Christoph Schnörr

Image and Pattern Analysis Group, Heidelberg University

GCPR 2015

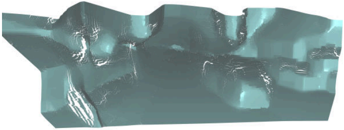
Motivation

- *monocular* reconstruction
  - \+ low-cost sensor
  - \+ no calibration
  - – pose needs to be estimated
  - – less beneficial parallax
- from *two frames*
  - \+ fast response
- *piecewise planar* structure
  - \+ suitable for urban scenes
- in a *continuous* framework
  - \+ does not require discrete plane candidates



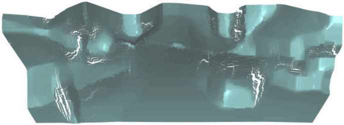
Motivation

- *monocular* reconstruction
  - + low-cost sensor
  - + no calibration
  - – pose needs to be estimated
  - – less beneficial parallax
- from *two frames*
  - + fast response
- *piecewise planar* structure
  - + suitable for urban scenes
- in a *continuous* framework
  - + does not require discrete plane candidates



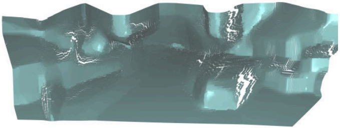
Motivation

- *monocular* reconstruction
  - \+ low-cost sensor
  - \+ no calibration
  - – pose needs to be estimated
  - – less beneficial parallax
- from *two frames*
  - \+ fast response
- *piecewise planar* structure
  - \+ suitable for urban scenes
- in a *continuous* framework
  - \+ does not require discrete plane candidates



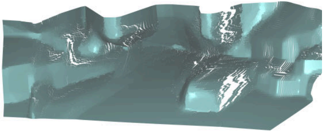
Motivation

- *monocular* reconstruction
  - \+ low-cost sensor
  - \+ no calibration
  - – pose needs to be estimated
  - – less beneficial parallax
- from *two frames*
  - \+ fast response
- *piecewise planar* structure
  - \+ suitable for urban scenes
- in a *continuous* framework
  - \+ does not require discrete plane candidates



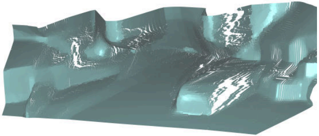
Motivation

- *monocular* reconstruction
  - + low-cost sensor
  - + no calibration
  - – pose needs to be estimated
  - – less beneficial parallax
- from *two frames*
  - + fast response
- *piecewise planar* structure
  - + suitable for urban scenes
- in a *continuous* framework
  - + does not require discrete plane candidates



Motivation

- *monocular* reconstruction
  - \+ low-cost sensor
  - \+ no calibration
  - – pose needs to be estimated
  - – less beneficial parallax
- from *two frames*
  - \+ fast response
- *piecewise planar* structure
  - \+ suitable for urban scenes
- in a *continuous* framework
  - \+ does not require discrete plane candidates



Related Work

F. Becker, F. Lenzen, J. H. Kappes, and C. Schnörr. Variational Recursive Joint Estimation of Dense Scene Structure and Camera Motion from Monocular High Speed Traffic Sequences. *Int J Comput Vision*, 105 (3):269–297, 2013.

C. Vogel, S. Roth, and K. Schindler. An Evaluation of Data Costs for Optical Flow. In *German Conference on Pattern Recognition (GCPR)*, 2013.

K. Yamaguchi, D. A. McAllester, and R. Urtasun. Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. In *ECCV*, 2014.

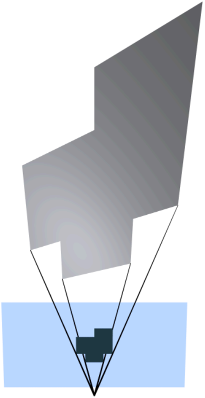
Problem Statement

Input

- optical flow, computed with DataFlow by Vogel et al. [2013]

Output

- planes $\{v_i \in \mathbb{R}^3\}$, $v^\top \tilde{x} = 1$
- on superpixels $\{\Omega_i \subset \Omega \mid i \in [1, n]\}$
- camera movement (up to scale)
  - $R \in \mathrm{SO}(3)$
  - $t \in S^2$



Workflow

$\hat{u}$

$\hat{u}^{-1}$

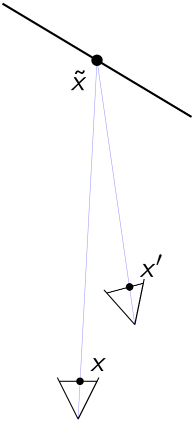
$$\boxed{\min_{R,t,v} E_u(R,t,v) + E_{\text{reg}}(v)}$$

The data energy reads

$$E_u(R,t,v) = \sum_{i=1}^{n} \sum_{x \in \Omega_i} w_{\hat{u}} \| u(x; R, t, v^i) - \hat{u}(x) \|_2^2$$

$$w_{\hat{u}} = \exp\left(- \frac{\| x - (\hat{u}^{-1} \circ \hat{u})(x) \|_2^2}{2\sigma_{\hat{u}}^2} \right)$$

Motion Field from Scene Parameters



- camera intrinsics are known
- apply camera motion

$$\tilde{x}' = R^\top(\tilde{x} - t)$$

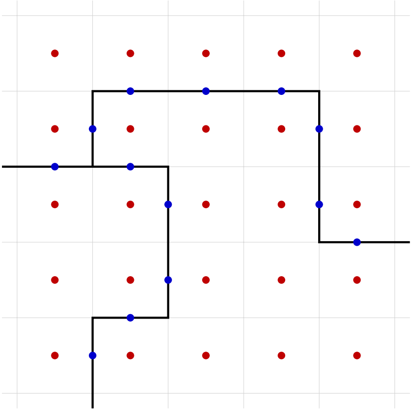
- project onto image plane,

$$x' = \pi(R^\top(\tilde{x} - t)), \quad \pi(x) = \frac{1}{x_3}x$$

- using $\tilde{x} = d(x)x$ and $v^\top x = \frac{1}{d(x)}$

$$u(x; R, t, v) = \pi(R^\top(I_3 - tv^\top)x) - x$$

Superpixel Domain



- $\mathcal{N}_\Omega$ contains neighboring superpixel pairs
- ∂^{ij} border between superpixels i and j
- x_c^i center (average) point of superpixel i
- $z(x)$ denotes inverse depth

Regularity

$$E_{\text{reg}}(v) = \lambda_z E_z(v) + \lambda_v E_v(v) + \lambda_p E_p(v)$$

- E_z smooths depth
- E_v smooths plane parameters
- E_p enforces positive depth
  - a soft hinge-loss function is applied to $z(x_c^i)$, the inverse depth at superpixel center

Piecewise Planar Regularization

- penalize differences of inverse depth on superpixel edges

$$E_z(v) = \sum_{(i,j)\in\mathcal{N}_\Omega} \sum_{x\in\partial^{ij}} \rho_C(\underbrace{x^\top v^i}_{z(v^i;x)} - \underbrace{x^\top v^j}_{z(v^j;x)})^2$$

- penalize jumps of plane parameters

$$E_v(v) = \sum_{(i,j)\in\mathcal{N}_\Omega} \rho_C(v^i - v^j)^2$$

Piecewise Planar Regularization

- penalize differences of inverse depth on superpixel edges

$$E_z(v) = \sum_{(i,j)\in\mathcal{N}_\Omega} \sum_{x\in\partial^{ij}} \rho_C(\underbrace{x^\top v^i}_{z(v^i;x)} - \underbrace{x^\top v^j}_{z(v^j;x)})^2$$

- penalize jumps of plane parameters

$$E_v(v) = \sum_{(i,j)\in\mathcal{N}_\Omega} \rho_C(v^i - v^j)^2$$

- choose the robust pseudo-Huber or Charbonnier function

$$\rho_C(x) = (x^2 + \epsilon)^\alpha - \epsilon^\alpha$$

with $\alpha = \frac{1}{4}$ (approximation to $L1$ norm)

Optimization

Overall energy function

$$E(X) = \|F(X)\|_2^2, \quad X = (R, t, v),$$

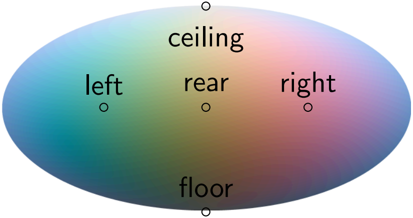
Optimization: *Levenberg-Marquardt* algorithm.

$$X^{k+1} = X^k + (J^\top J + \mu^k I)^{-1}(J^\top F), \quad J = \frac{\partial F}{\partial X}$$

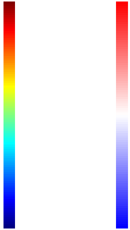
- for the rotation, we have $R^{k+1} = R^k \exp(\omega)$, $\omega \in \mathfrak{se}(3)$
- t is projected onto the sphere after each update
- v is not constrained

Evaluation

- we compare against SPS-St by Yamaguchi et al. [2014] on the KITTI *stereo/flow* dataset
- a quantitative evaluation as in Becker et al. [2013] is done in the paper
- we provide scenes with reference normal information
- the following color representations are used



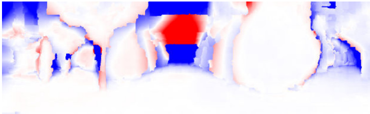
LAB plane encoding



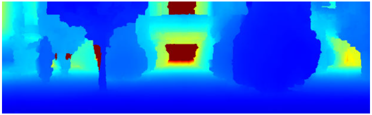
depth Δdepth

KITTI Stereo/Flow Sequence 9

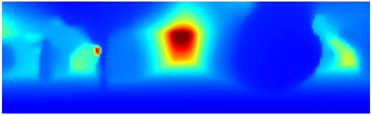
reference frame



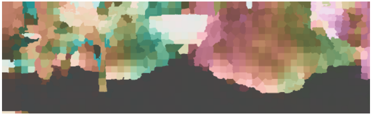
depth difference



depth (SPS-St, stereo)



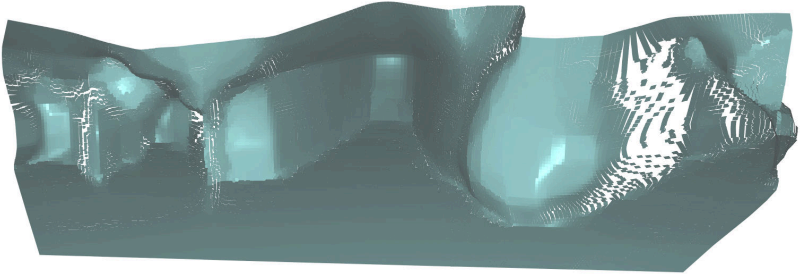
depth (ours, monocular)



normals (SPS-St)

normals (ours)

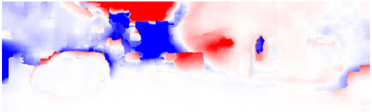
KITTI Stereo/Flow Sequence 9



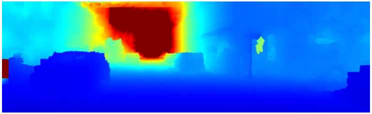
rendered reconstruction

KITTI Stereo/Flow Sequence 19

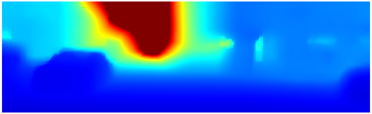
reference frame



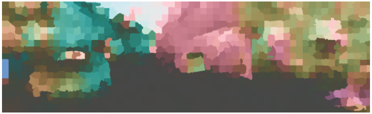
depth difference



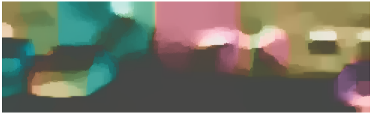
depth (SPS-St, stereo)



depth (ours, monocular)

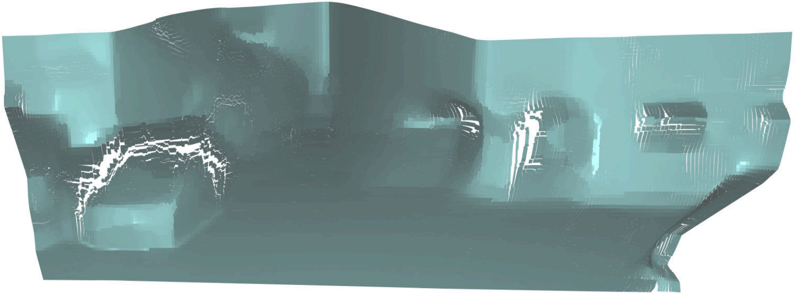


normals (SPS-St)



normals (ours)

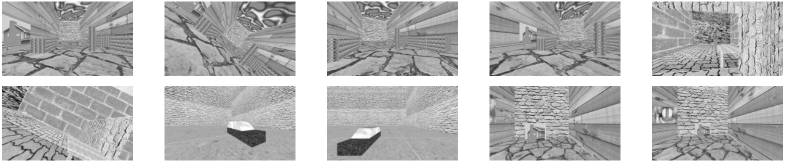
KITTI Stereo/Flow Sequence 19



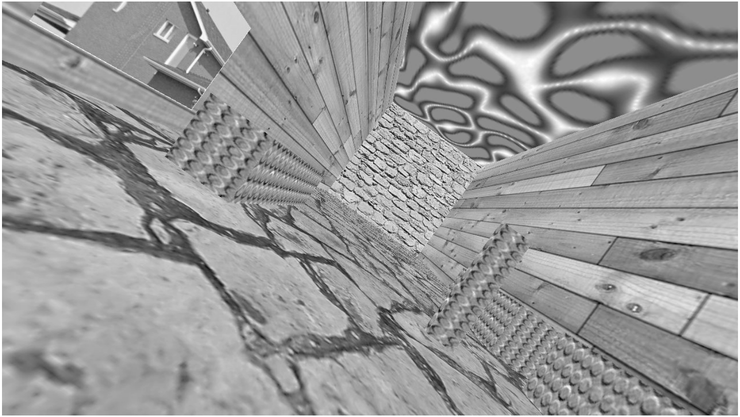
rendered reconstruction

Rendered Scenes

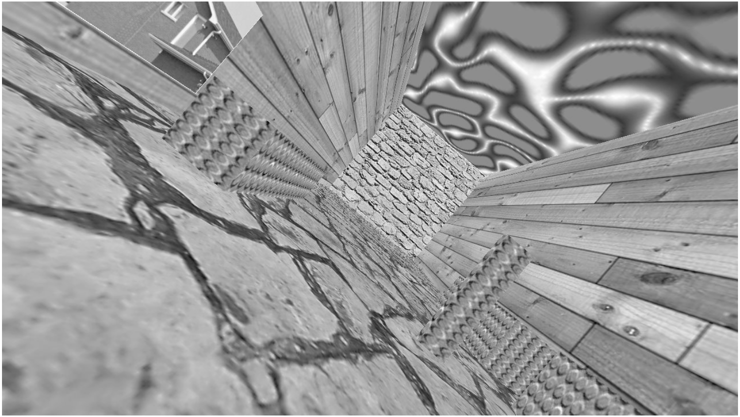
- 10 sequences
- 25 frames per sequence
- ground truth depth and normals



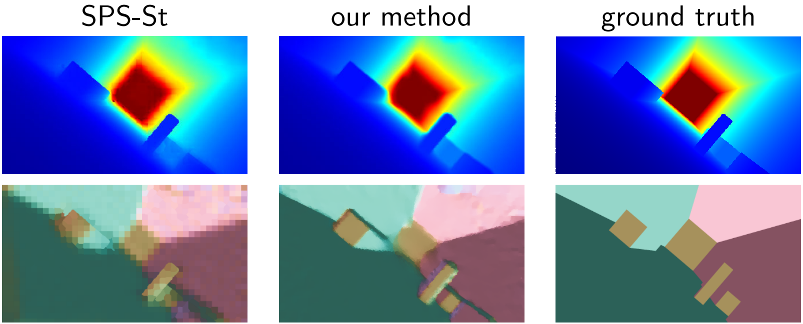
Scene 2, Frame 9



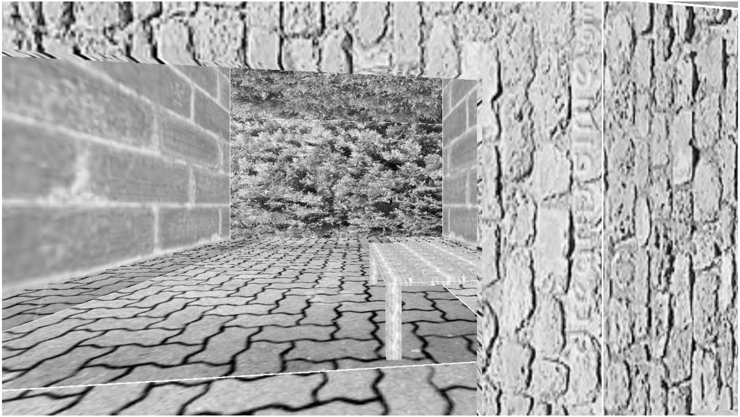
Scene 2, Frame 10



Scene 2, Frame 10 Results

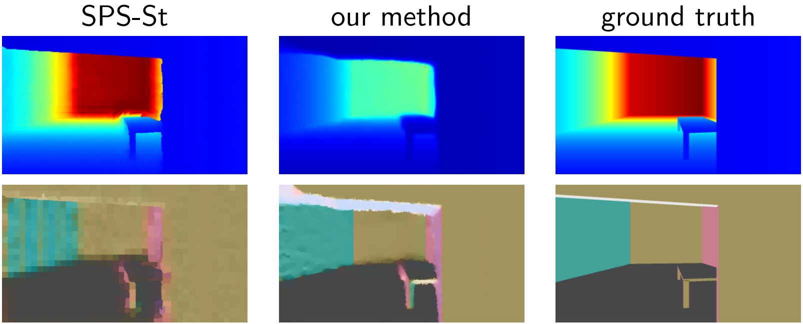


Scene 5, Frame 9

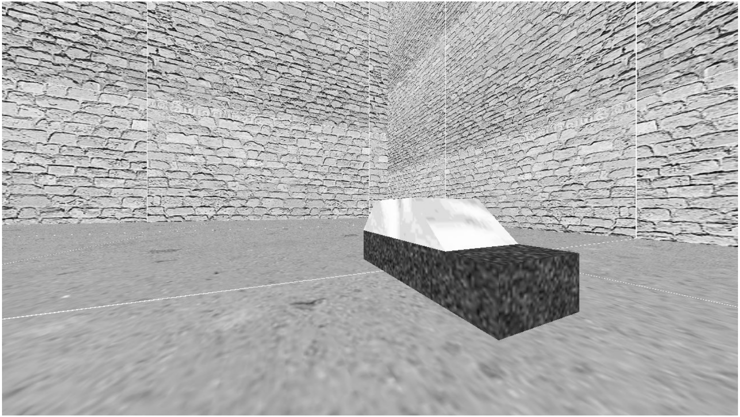


Scene 5, Frame 10

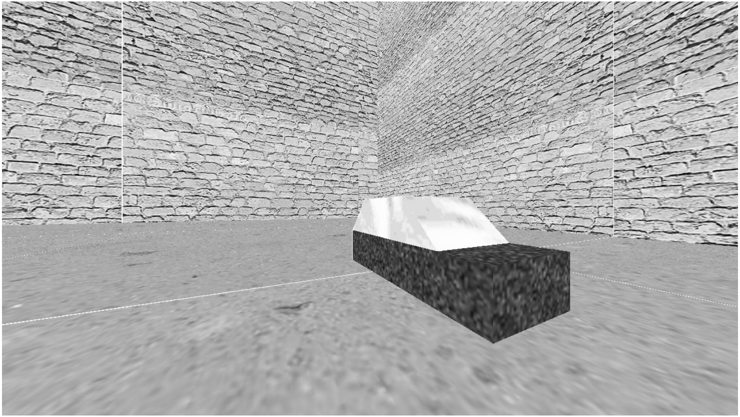
Scene 5, Frame 10 Results



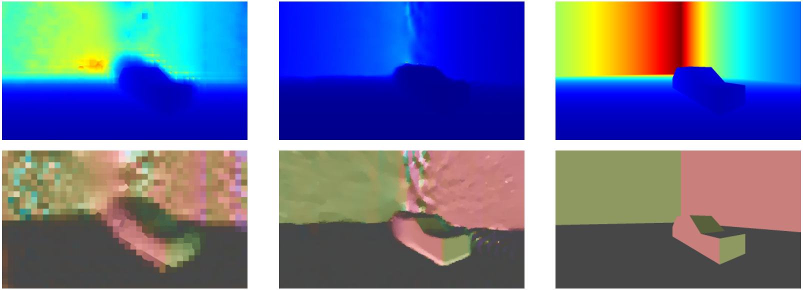
Scene 2, Frame 9



Scene 2, Frame 10



Scene 7, Frame 10 Results



Normal Reconstruction Error

	mean [deg.]	$p_{1\mathrm{deg.}}$ [%]	$p_{5\mathrm{deg.}}$ [%]	$p_{10\mathrm{deg.}}$ [%]
SPS-St	14.8	79.4	46.4	33.4
our method	11.5	58.4	31.1	22.7

Conclusion and Future Work

- we presented a two-frame monocular reconstruction method
- both camera movement and scene are unknown
- scene reconstruction accuracy is similar to state of the art stereo approaches
- method can be extended to image sequences
- operate on image gray values rather than optical flow
- detection of dynamic objects

Estimating Vehicle Ego-Motion and Piecewise Planar Scene Structure from Optical Flow in a Continuous Framework

Andreas Neufeld, Johannes Berger, Florian Becker, Frank Lenzen and Christoph Schnörr

Image and Pattern Analysis Group, Heidelberg University

GCPR 2015