TOSHIBA
Leading Innovation

UNIVERSITY OF OXFORD

Coarse-to-fine Planar Regularization for Dense Monocular Depth Estimation

Stephan Liwicki Christopher Zach Ondrej Miksik Philip H. S. Torr

1 Introduction

Simultaneous localization and mapping (SLAM) using the whole image data is an appealing framework to address shortcoming of sparse feature-based methods – in particular frequent failures in textureless environments. Hence, direct methods bypassing the need of feature extraction and matching became recently popular. Many of these methods operate by alternating between pose estimation and computing (semi-)dense depth maps, and are therefore not fully exploiting the advantages of joint optimization with respect to depth and pose.

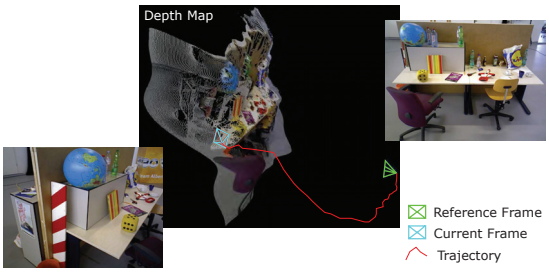

Figure 1: During keyframe-to-frame comparison a dense depth map is build. An incremental implementation allows for longer tracks with a single reference frame in real-time.

Contributions

- Global energy for planar inverse depth that is optimized iteratively
- Coarse-to-fine strategy that refines depth and pose truly simultaneously
- Semi-dense version, computing depth twice as fast as LSD-SLAM [1] on CPU
- Quantitative pose and depth evaluation on the TUM dataset [4]

2 Global Energy

The global energy (using a single reference image) is formulated as

$$E_{All}(\mathcal{S}, \Xi) = \sum_{t=1}^{T} E^{(t)}_{Match}(\mathcal{S}, \xi_t) + E_{Smooth}(\mathcal{S}), \tag{1}$$

- Photometric error $E^{(t)}_{Match}$
- Spatial smoothing E_{Smooth} (over inverse depth values)
- Camera poses $\Xi = (\xi_t)_{t=1}^{T}$
- Planes $\mathcal{S} = (\mathbf{s}_i^{\mathrm{T}})_{i=1}^{|\mathcal{X}|}$ (over-parametrization of $\mathcal{D} = (d_i)_{i=1}^{|\mathcal{X}|}$, where $d_i = \mathbf{s}_i^{\mathrm{T}}\mathbf{x}_i$)

Temporal Smoothing

We modify E_{All} to be suitable for an incremental approach for current pose ξ_T:

$$E_{All}(\mathcal{S}, \xi_T) = E^{(T)}_{History}(\mathcal{S}) + E^{(T)}_{Match}(\mathcal{S}, \xi_T) + E_{Smooth}(\mathcal{S}). \tag{2}$$

- Second order approximation of $E^{(T)}_{History}$ provides temporal smoothness

Photometric Matching Cost

We express the photometric error by

$$E^{(T)}_{Match}(\mathcal{S}, \xi_T) = \sum_{\mathbf{x}_i \in \mathcal{X}} \| I(\mathbf{x}_i) - I_T\left(W(\mathbf{x}_i, d_i, \xi_T)\right) \|_{\tau_{Match}}. \tag{3}$$

- Warping $\mathbf{x}'_i = W(\mathbf{x}_i, d_i, \xi_t) = \hom\left(\mathbf{R}_t^{\mathrm{T}}\left(\mathbf{x}_i - \mathbf{t}_t d_i\right)\right) = \hom\left(\mathbf{R}_t^{\mathrm{T}}\left(\mathbf{x} - \mathbf{t}_t \mathbf{s}_i^{\mathrm{T}}\mathbf{x}\right)\right)$
- Robust smooth truncated quadratic $\| \cdot \|_{\tau_{Match}}$ given in [2]

Local Spatial Plane Regularizer

The smoothness constraint $E_{Smooth}(\mathcal{S})$ is expressed in inverse depth:

$$E_{Smooth}(\mathcal{S}) = \lambda_{Smooth} \sum_{\mathbf{x}_i \in \mathcal{X}} \sum_{\mathbf{x}_j \in \mathcal{N}_i} \| \mathbf{s}_i^{\mathrm{T}}\mathbf{x}_i - \mathbf{s}_j^{\mathrm{T}}\mathbf{x}_i \|_{\tau_{Smooth}}. \tag{4}$$

- Balancing term λ_{Smooth}
- Smoothness assumption inspired by stereo setups

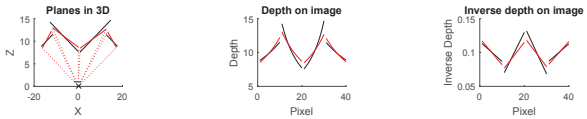

Figure 2: Top view of planes. Pixels in 3D space are aligned via smoothing in the inverse depth image.

[1] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *ECCV'14*, pages 834 – 849, 2014.
[2] H. Li, R. Summer, and M. Pauly. Global Correspondence Optimization for Non-Rigid Registration of Depth Scans. *European Symp. Geometry Processing*, 27(5):1421–1430, 2008.
[3] H. Mobahi and J. Fisher. On the Link between Gaussian Homotopy Continuation and Convex Envelopes. In *Int. Conf. Energy Minimization Methethods Computer Vision and Pattern Recognition, EMMCVPR'15*, pages 43–56, 2015.
[4] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *IROS'12*, 2012.

3 Optimization Strategy

We find pose as well as depth though Levenberg-Marquardt optimization.

- Locally linarize I_T (as in KLT)
- Graduated optimization [3]
  - scale-space pyramid
  - proposed restricted depth

Figure 3: We incrementally increase depth resolution.

Input: Keyframe I and images $(I_t)_{t=1}^{T}$.
Output: Final pose ξ and depth hypothesis $\mathcal{S}$.

$\mathbf{s}_i \leftarrow [0\,0\,1]^{\mathrm{T}}$ and initial depth certainty $\Lambda_i \leftarrow 0$ for all $\mathbf{x}_i \in \mathcal{X}$.
compute resolution pyramid for the keyframe I.
$\xi \leftarrow (\mathbf{I} \in \mathbb{R}^{3\times3}, [0\,0\,0]^{\mathrm{T}})$
for each frame I_t **do**
  compute resolution pyramid for the frame I_t.
  for each pyramid level **do**
    optimize ξ via lie algebra $\mathfrak{se}(3)$ through Levenberg-Marquardt.
    repeat
      update ξ (and $\mathbf{s}_i \leftarrow \mathbf{s}_i + \mathbb{I}_c(\mathbf{x}_i)\Delta_c$ if applicable).
      introduce new component Δ_c.
      estimate $\mathbb{I}_c(\mathbf{x}_i)$ via eigenvector of $\sum_{\mathbf{x}_i \in \mathcal{X}} \nabla_{\mathbf{s}_i} \nabla_{\mathbf{s}_i}^{\mathrm{T}}$.
      optimize ξ and Δ_c through Levenberg-Marquardt.
    until improvement below $\epsilon_{Complex}$ or maximum C reached
  end for
  update precision Λ_i and depth $\mathbf{s}_i^*$ for temporal constraint.
end for

Algorithm 1: Dense Incremental Planar Depth Estimation

Coarse-to-fine Depth Perception

We enforce low complexity in the update of $\mathcal{S}$:

$$\mathbf{s}_i = \mathbf{s}_i^* + \sum_{c=1}^{C} \mathbb{I}_c(\mathbf{x}_i)\Delta_c, \tag{5}$$

- Indicator function $\mathbb{I}_c : \mathcal{X} \to \{+1, -1\}$ (updates are constrained to 2^C values)
- Greedily add one component Δ_c at a time (3 (Δ_c) + 6 (SE(3)) parameters)

Advantages in Optimization Conditions

- Regularization across all pixels is enforced for the update of planes
- The image planes encode inverse depth hierarchically from coarse to fine
- The incremental algorithm enables fast, joint optimization of pose and depth
- Depth updates are simple, but yield rich depth maps after repeated updates

4 Evaluation

We evaluate our methods, dense DIP and semi-dense SIP, on 6 TUM and 7 own videos.

Quantitative Results

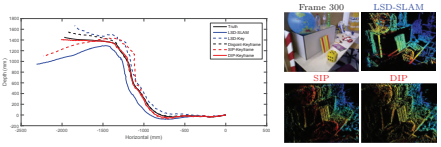

Figure 4: Quantitative trajectory (left) and its depth (right) of an example sequence.

- DIP (one keyframe) performs similar to LSD-SLAM (many keyframes) for pose
- DIP benefits from larger baselines for depth estimation
- LSD-Key (LSD-SLAM with single reference frame) is less favourable
- Our semi-dense SIP performs well for small motion
- The disjoint (alternating) version is worse in virtually all experiments

Qualitative Results

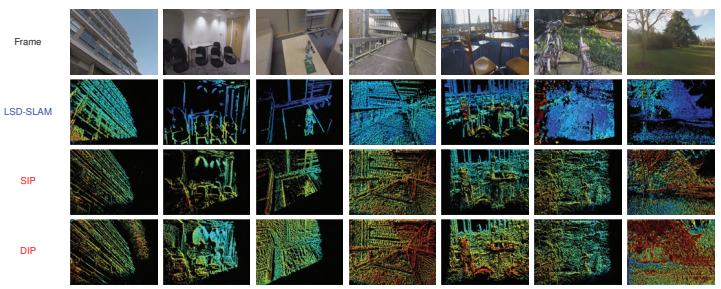

Figure 5: Qualitative depth results for our videos.

- SIP, DIP produce more globally consistent depths in contrast to LSD-SLAM
- Even in non-planar scenes, our method produces sensible results
- Our method converges to local minima if initial $\mathbf{s}_i \leftarrow [0\,0\,1]^{\mathrm{T}}$ is significalty wrong

Running time

- DIP is highly parallel and computs in real-time on GPU
- SIP estimates depth *and* pose twice as fast as LSD-SLAM using CPU (30 fps)