UnDeepVO: Monocular Visual Odometry through Unsupervised Deep Learning

Bolaños T.
Kurki L.
Rehn A.
Zaka A.
Zhao Z.

October 16, 2020

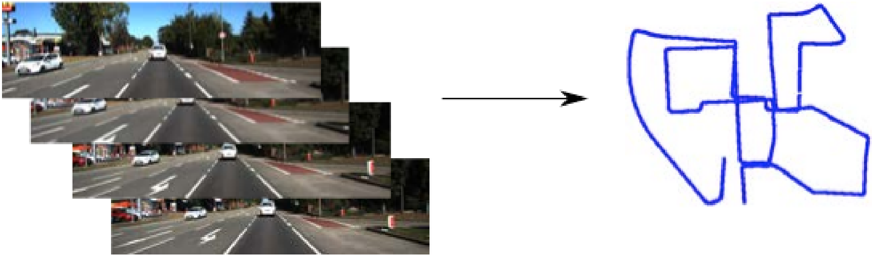
Outline

- A monocular visual odometry system
- Paper by Ruihao Li, Seng Wang, Zhiqiang Long and Dongbing Gu

Introduction
Visual odometry

- Goal
  - Robot localization using only visual information

- Goal
  - Use consecutive monocular images to construct a path of robot movement

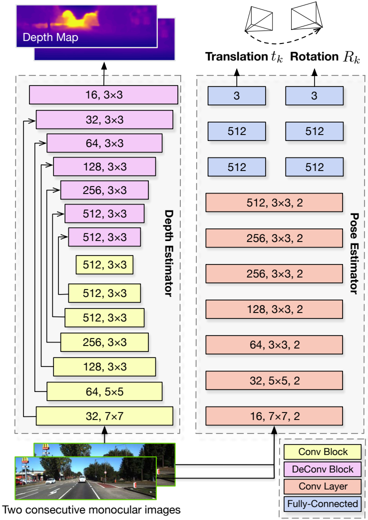


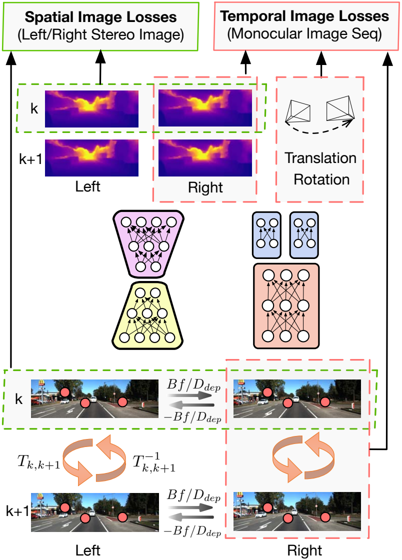
- Supervised Learning
  - CNN for 6-DOF pose regression
  - Video clips
  - Optical flow
  - DeMoN
  - Visual inertial odometry
  - 'Spatial transformer'
- Unsupervised Learning
  - Photometric constraint of stereo imaging
  - Consecutive monocular Imaging

- Monocular stereo imaging based VO system
- Based on deep learning
- Unsupervised
    - No need for labeled training data
- Pose estimation
- Depth estimation
- Absolute scale retrieval
- Evaluation using KITTI dataset

System Overview

Architecture





Spatial Image Losses
(Left/Right Stereo Image)
Temporal Image Losses
(Monocular Image Seq)
k
k+1
Left
Right
Translation
Rotation
k
Bf/D_{dep}
$-Bf/D_{dep}$
$T_{k,k+1}$
$T^{-1}_{k,k+1}$
k+1
Bf/D_{dep}
$-Bf/D_{dep}$
Left
Right

Objective Losses

Spatial Losses

The spatial losses are based on the fact that, given the structure of stereo cameras, for a pixel $p_l(u_l, v_l)$ on the left image and $p_r(u_r, v_r)$ on the left image:

$$u_l = u_r \quad \text{and} \quad v_l = v_r + D_p$$

- Photometric Consistency Loss (Image reconstruction)

$$L_{pho} = \lambda_s L^{SSIM}(I, I') + (1 - \lambda_s) L^{l_1}(I, I')$$

- Disparity Consistency Loss (Depth)

$$L_{dis} = L^{l_1}(D_{dis}, D'_{dis})$$

- Pose Consistency Loss (Camera orientation)

$$L_{pos} = \lambda_p L^{l_1}(t_l, t_r) + \lambda_o L^{l_1}(R_l, R_r)$$

This is based on the reconstruction of pixels on time k and $(k+1)$ as

$$p_{k+1} = KT_{k,k+1}D_{dep}K^{-1}p_k$$

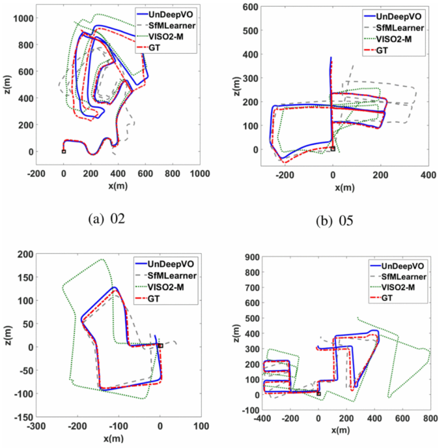
- Photometric Consistency Loss (Image reconstruction)

$$L_{pho} = \lambda_s L^{SSIM}(I, I') + (1 - \lambda_s)L^{l_1}(I, I')$$

- 3D Geometric Registration Loss (Adding depth with $P(x, y, z)$)

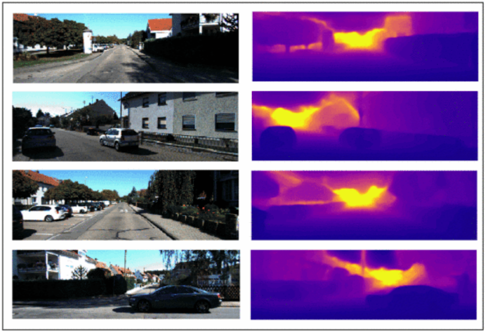
$$L_{geo} = L^{l_1}(P, P')$$

- KITTI Odometry Dataset
- Comparison between UnDeepVO, SfMLearner VISO2-M and ORB-SLAM-M
- UnDeepVO qualitatively closest to the ground truth for all sequences



(a) 02 (b) 05 (c) 07 (d) 08

- UnDeepVO also produces a scaled depth map
- Depth of objects estimated accurately
- Model outperforms some competitors but not all
  - Only part of KITTI dataset used
  - Lower image resolution
  - Temporal image sequence loss might have introduced noise



Conclusions

- First unsupervised Visual Odometry model
  - Trained with unlabeled stereo images
  - Uses stereo image pairs to recover the scale
    - Scale can not be recovered from monocular images
- Performs inference on monocular images
- Pose and dense estimations for recovering the trajectory
  - One CNN for depth estimation
  - Another CNN for pose estimation
- Outperforms previous methods in almost all cases
- Plans to extend to a full SLAM system

Contributors

- Bolaños Tlahui
  - Objective Losses
- Kurki Lauri
  - Evaluation
- Rehn Aki
  - Organization, introduction, conclusions
- Zaka Ayesha
  - UnDeep VO Key Contributions
- Zhao Zhao
  - System Overview