Deep Relative Pose Estimation for Stereo Camera

3D Vision Project Proposal
Supervised by: Zhaopeng Cui
March 9, 2018

GROUP MEMBERS

Noah Isaak

Max Huerlimann

I. DESCRIPTION OF THE PROJECT

This project aims at designing an end-to-end neural network which can can estimate the depth and relative pose between consecutive frames of a stereo camera. It is inspired by the work done with an unsupervised network for single-view monocular cameras [3]. Here the depth is directly learned from a single image, and thus quite noisy and lacks of generalization to unknown scenes. With a pair of stereo images, we can get more accurate depth information, which should be able to further improve the pose estimation. What's more, it has better generalization. Given the two pairs of stereo images as input, our network will contain two parts: depth estimation and pose regression. We will adopt some existing work for depth estimation, and mainly focus on the design of the pose estimation. The depth estimation has been shown to work well with a stereo configuration [2]. For training the KITTI dataset will be used, which provides scenes captured with stereo cameras. If the time allows, there will be attempted to fit an explanatory mask as well.

II. WORK PACKAGES AND TIMELINE

To get familiar with the recent work, at first familiarization with the state-of-the-art algorithms will take place. The Dispnet method, developed by researchers at the university of Freiburg [1], will be implemented on our computers and played around with. Then we will adapted this deep stereo method as the first part of our pipeline. After that, we will focus on the design of the subnetwork for pose estimation by adapting the network done by Zhou [3]. The network will be trained through the KITTI dataset (and possibly others), preferably in an unsupervised manner. If performance is unsatisfactory, supervised learning will be experimented with. The last step will be to validate the results on available datasets. If the time allows, the topic of semantics will be tackled. An explanatory mask could be learned simultaneously to differentiate the moving objects and improve the robustness of the method. All the code will be implemented through Python on a PC and the TensorFlow framework will be used. The intended timetable can be seen in table I.

TABLE I: Timetable

Workpackage	**Timeframe**	**Supervised by**
Getting familiar with existing code	12/03 - 17/03	Max & Noah
Adaptation of Dispnet	18/03 - 21/03	Noah
Design of pose estimation network	21/03 - 07/04	Max
Training of the network on KITTI dataset	08/04 - 01/05	Max & Noah
Validation of network	02/05 - 15/05	Max & Noah

III. Outcomes and Demonstration

The main expected outcome is an improvement of depth perception with respect to the single view monocular algorithm. This could lead to a general improving of relative pose estimation and possibly semantics. This can be demonstrated on available datasets or possibly live, if there is a stereo camera available. To be able to measure improvement, either the existing code from [3] can be used to create reference data from new datasets or existing data from the paper itself can be used as reference.

References

[1] N.Mayer, E.Ilg, P.Häusser, P.Fischer, D.Cremers, A.Dosovitskiy, and T.Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[2] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. *CoRR*, abs/1612.02401, 2016.

[3] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.