Monocular Visual Odometry

alex.kreimer

February 10, 2017

We propose to build upon our experience with the stereoscopic VO to build a monocular solution. Traditionally, monocular VO is considered a more complex problem, e.g., it is harder to achieve the same result with a single camera. Below we provide a short overview of modern VO algorithms with an outline of the system we propose to build.

Modern VO algorithms consist of the following stages: 1. Feature correspondence across a subset of selected frames (keyframes). The choice of feature detector/descriptor with a matching procedure is a tradeoff of accuracy/runtime. 2. Keyframes selection. The complexity of the algorithm grows with a number of keyframes. It is important to implement a sensible keyframes selection policy that would provide a well spread set of frames with significant parallax and plenty of loop closure matches. 3. An initial estimate of the keyframe poses as an input to the non-linear optimization. 4. A local map where the optimization is focused to achieve scalability 5. The ability to perform fast global optimizations (e.g. pose graph) to close loops in real time.

Since 2012, the Computer Vision field has undergone the deep learning revolution.

Visual Odometry is one of the fields, that is yet to be conquered by the neural nets. We propose to conduct series of experiments to solve the monocular visual odometry using the neural nets.

The idea is to model the visual odometry as a regression problem and to use the convolutional neural nets as a regressor (see e.g., [4]). This way, the feature extraction and the motion estimation problem (roughly corresponds to the steps 1-3 above) will be solved by the end-to-end trained convolutional net. Robot state will be modeled with the recurrent neural networks (e.g. LSTM), see e.g., [1]. We will experiment with different architectures, training schemes, etc.

We will use a common Caffe [3] framework to perform the experiments (see [2] for the LSTM implementation).

Note, while the process seems to be straightforward, the field evidence suggests that it is challenging to obtain good results. It will probably be impossible to compete with the state-of-the art results. On the other hand, the machine learning approaches has a much broader horizon than the current state of the art methods, which may justify their use.

Suggested roadmap:

- Literature survey - 1 month.
- Train a CNN to solve the visual odometry without temporal information (e.g., no LSTM): 5 months.
- Train LSTM network to improve the single frame CNN result: 6 months.

References

[1] Ronald Clark et al. "VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem". In: *CoRR* abs/1701.08376 (2017). URL: http://arxiv.org/abs/1701.08376.

[2] Jeff Donahue et al. "Long-term Recurrent Convolutional Networks for Visual Recognition and Description". In: *CoRR* abs/1411.4389 (2014). URL: http://arxiv.org/abs/1411.4389.

[3] Yangqing Jia et al. "Caffe: Convolutional Architecture for Fast Feature Embedding". In: *CoRR* abs/1408.5093 (2014). URL: http://arxiv.org/abs/1408.5093.

[4] Vikram Mohanty et al. "DeepVO: A Deep Learning approach for Monocular Visual Odometry". In: *CoRR* abs/1611.06069 (2016). URL: http://arxiv.org/abs/1611.06069.