Review for Learning Monocular Visual Odometry via Self-Supervised Long-Term Modeling

Irem Arpag

March 29, 2021

The paper deals with a long-standing problem, visual odometry, estimating the ego-motion incrementally by using visual input, that becomes more challenging with the necessity of consistency over long sequences which means a video level performance is required. The researchers in the article propose a monocular VO system in a self-supervised manner based on a significant long-term modelling to overcome the problem by combining both traditional geometric and novel learning-based methods. With a stage-wise training strategy, in the first stage of it, by training a depth network and a pose network with two layer convolutional LSTM with short snippets and in the second one, by training only a light-weight second layer convolutional LSTM, they pre-extract features using the first stage by discarding the heavy feature extraction layers that allows with long sequences around 100 frames in training time, to be thought as the first deep-learning approach for VO that takes long sequences as input in the training stage. The cycle consistency constraints between two-layer LSTM estimations also serves as a mini loop closure to strengthen the transitivity consistency of poses. Exploiting two-layer LSTM (long short-term memory of RNN) network for pose estimation in the proposed method s relating with sequential modeling which is based on recurrent neural network (RNN) and VO as the estimation of full trajectory over a long sequence of frames is a sequential problem and modelled with RNN. The basic components of the model that are network and training scheme are completely inspired by a well-studied geometric module and this significant design produced better results than the existing self-supervised baseline.

When comparing self-supervised manner of the model with fully supervised methods, while the said model needs only monocular video frames, the others require large-scale datasets with ground-truth annotations which are very sparse and cause labor and time consuming despite their good performance. Additionally, contrast to other self-supervised methods, the performance of ego-motion estimation of which are worse than traditional VO methods although they achieve good performance on single view depth estimation but not more than 5-frame snippets, that not enough for long information, the said self-supervised model directly optimizes over long sequences via long-term modelling that highlights it most.

The proposed model has some limitations, the most noticeable one is that the progress on the rotation prediction is not good enough as the translation since it accomplishes a good camera pose estimation performance in terms of translation error. The bias within the dataset is thought to be the potential reason of the large rotation error and the solution is stated to train the model with more various video sequence and synthetic data. The failure of system under over-exposure scenarios because of relying on visual input to extract information is another weakness of the model.

The KITTI Odometry Dataset that contains 22 sequences are used for evaluation but only the sequences 00-10 have ground truth trajectory labels so sequences 00-08 are used for training purposes and the last two (09-10) for validation ORB-SLAM2 is also used for stereo version since the ground truth labels of sequences 11-21 of KITTI Odometry Dataset are not available to get predictions as pseudo. For indoor environments TUMRGB-D Dataset is exploited. As ablation study, different variations of method is evaluated by using sequences 09 and 10of the KITTI Odometry Dataset. By adding more components, they test the performance of the model as follows; 1) Baseline two consecutive frame, 2) One-layer conv. LSTM, 3) Two-layer conv. LSTM, 4) Two-layer conv. LSTM + Two stage training (final model). The qualitative and quantitative results confirm that adding each of the components gradually improves the overall performance Number 4 variation (final model) achieves a state-of-art accuracy within self-supervised methods. The final model is also compared with some other supervised and self-supervised methods by a large margin and among supervised methods the results are similar except Beyond Tracking's performance on Seq 10. For indoor environments, a challenging dataset TUMRGB-D is used, and the proposed model is compared with some significant traditional and learning-based methods. It is evaluated that traditional geometric methods perform best on some sequences but on some others, they fell to produce results because of tracking failure. The said self-supervised model could not perform better results than supervised models that is thought to be related with limited amount of training data.

With its two-layer conv. LSTM module that provides of modeling the long-term dependency in the pose estimation and with its stage-wise training strategy which allows the network to see beyond short snippets, the system prevails state-of-art performance among self-supervised methods. Detecting loops and performing full loop-closure is the new challenge of the researchers as future plan since they do not have any mechanism in the existent form. The current model also inspired to some follow-up works, one of them is focusing on estimating depth and camera poses from a monocular video by combining two complimentary techniques [1].

REFERENCES

[1] Johannes Kopf, Xuejian Rong, Jia Bin Hung, Robust Consistent Video Depth Estimation, arXiv Dec 2020.