Dexter Watkins

Paper Summary

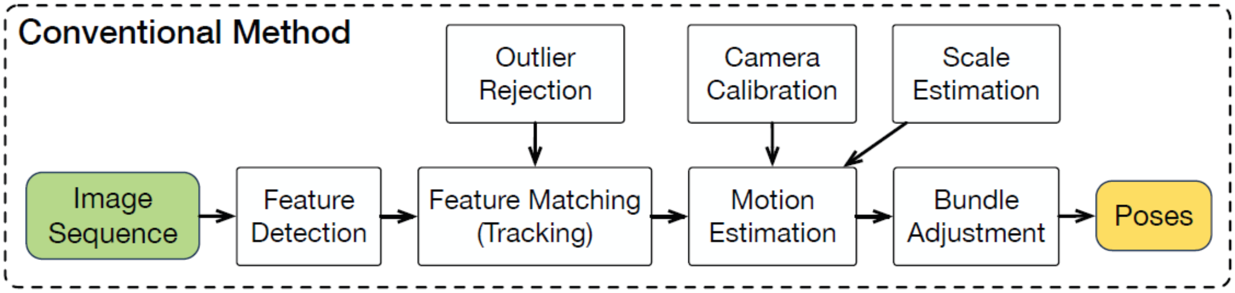
Wen, H., Wang, S., Clark, R., & Trigoni, N. (2017). {DeepVO}: Towards End to End Visual Odometry with Deep Recurrent Convolutional Neural Networks. *Icra*, 2043–2050. https://doi.org/10.1109/ICRA.2017.7989236

Overall Goal

The goal of this paper is to show that end-to-end deep learning techniques are a viable alternative to the traditional visual odometry processing pipeline. Additionally, the goal is to show that a deep learning visual odometry system requires little to no fine-tuning and does not depend on prior knowledge of a new environment.

Background

Visual Odometry (VO) is a method of extracting pose and location estimates for a mobile robot from an image sequence. Traditionally, this is done through a computer vision image processing pipeline that identifies features in an image frame, removes outliers, and tracks these features through sequential frames. Coupled with a camera calibration model and a dynamic model of the system, motion can then be estimated and integrated to determine a robot's most likely pose.



Issues with this "traditional" approach are that the dynamic model necessary for accurate pose estimation must be made by hand, a process that can be non-trivial for constrained, non-holonomic systems. Each processing module in the pipeline must be hand-tuned to a specific system by an individual with significant knowledge and training, and cannot be quickly modified to work with a new mobile system. This can include anything from modifications to image hardware to changes in the mobile base or method of locomotion.

Problems Addressed

The claim in this paper is that a Recurrent Convolutional Neural Network (RCNN) can be created that, given a proper training dataset, can achieve accurate visual odometry estimations with a significant decrease in the expertise needed to create such a system.

To be more specific, the implication is that by using an RCNN that only requires a properly made dataset with which to train, the issue of needing specific modules of the VO process to be hand-tuned by experts is obviated. There would then be no need for significant domain expertise to build a VO system.

More explicit problems that are stated by the authors are: 1) The development of an end-to-end VO system. 2) Development of a system able to be generalized to new environments, and 3) Implicit modeling of sequential dependences and complex motion dynamics of a mobile system.

Problem Solving Approach

To address these problems, a deep learning architecture was proposed that consists of 1) a Convolutional Neural Network based feature extraction that feeds into a 2) Recurrent Neural Network of LSTM for sequential pose estimation.

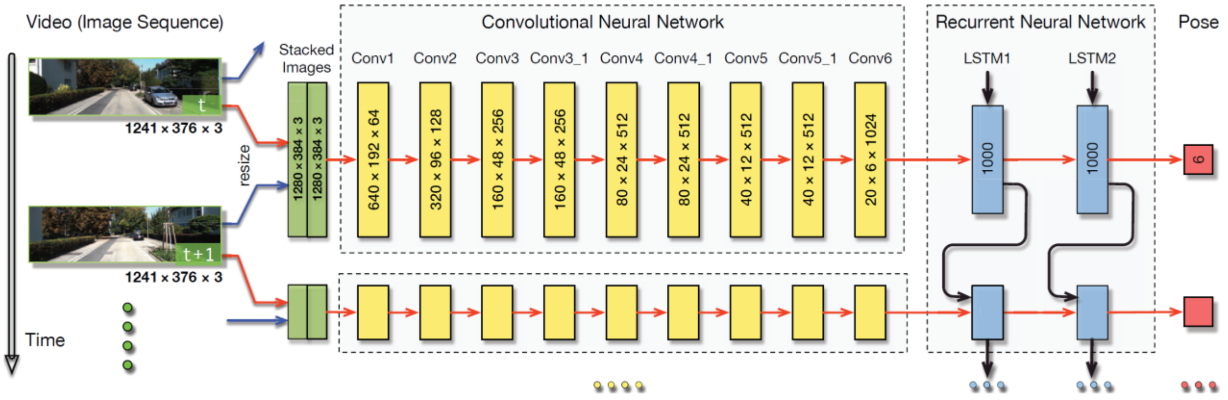


Fig. 2. Architecture of the proposed RCNN based monocular VO system. The dimensions of the tensors shown here are given as an example based on the image size of the KITTI dataset. The CNN ones should vary according to the size of the input image. Camera image credit: KITTI dataset.

The goal is that the CNN will be trained for identify features and the LSTM provides sequential memory so that pose estimation can be accurately determined. From this, using the well-formed and frequently used KITTI dataset, the model can be trained.

Results

Root Mean Squared Error of translational and rotation movement are used as metrics for overall performance, and shows that the DeepVO method is comparable to traditional approaches. Additionally, some cursory detail is given to the effects of model overfitting and the tweaking of model hyperparameters

Contributions

I believe that the main contribution of this paper is the validation of deep learning techniques in visual odometry. The system development in the paper did not achieve *better* results than traditional methods, but it did achieve *comparable* results at a *much lower cost* in time and expertise. It helps to pave the path for more rigorous techniques utilizing deep learning for Visual Odometry.