Deep SfM

Report (Work in Progress)

This page is meant to be a comprehensive collection of related literature, explanation of the problem, the approaches taken and the results obtained.

Abstract

Structure from Motion is a key problem in 3D computer vision which deals with reconstructing the 3D structure of the world seen by a camera from its 2D images. Humans can naturally perceive and understand the 3D world from the its 2D projection on the retina. In fact, humans are good at estimating the 3D structure by looking at 2D images taken from cameras. However, this is a very hard problem for computers.

In this work, we propose to explore how a generalizable SfM system can be built using deep neural networks by exploiting the visual priors learnt by them. We wish to build a model that predicts a 3D model of the world seen by a sequence of images and the camera poses that best explain the data.

Related Work

Fragkiadaki et al. (2017) propose a CNN based architecture for motion estimation in videos that decomposes frame-to-frame pixel motion in terms of scene and object, depth, camera motion and 3D object rotations and translations. They utilize geometric information to propose a self-supervised learning process. However the architecture only relies on pairs of frames to learn 3D structure and a separate processing step is required to obtain a globally consistent structure from all the predictions on a sequence of images.

Kendall, Grimes, and Cipolla (2015) and derived work Kendall and Cipolla (n.d.) Laskar et al. (2017) predict the camera pose from a single image after learning a map-like representation from training images. However, these approaches require ground truth poses and/or 3D structure.

Formal problem

Given only a sequence of images, how can we reconstruct a 3D model of the scene seen by the images and the relative camera poses that jointly explain the model and images. We focus on the following key aspects of the problem:

- General solution: The model should not learn a map from training images but instead learns to associate input images visually along with priors to build a 3D model. It is crucial that given an unseen sequence of images, the model should be able to predict a plausible explanation of the 3D structure and poses.
- Interpretability of the model: How can we translate the underlying representation learnt by the model to meaningful formats (depth, point cloud representation, object motion, etc)
- Geometry aware: The model must learn to exploit geometric structure in the images and be supervised by geometry based signals.
- Scalable: The approach must be indepent of the number of images and be able to globally reason from all the input and not rely on a set of local solutions followed by global optimization as post-processing.

References

Fragkiadaki, Aikaterini, Bryan Seybold, Rahul Sukthankar, Sudheendra Vijayanarasimhan, and Susanna Ricco. 2017. "Self-Supervised Learning of Structure and Motion from Video." In *Arxiv (2017)*. https://arxiv.org/abs/1704.07804.

Kendall, Alex, and Roberto Cipolla. n.d. "Geometric Loss Functions for Camera Pose Regression with Deep Learning." In.

Kendall, Alex, Matthew Grimes, and Roberto Cipolla. 2015. "Posenet: A Convolutional Network for Real-Time 6-Dof Camera Relocalization." In *Computer Vision (Iccv), 2015 Ieee International Conference on*, 2938–46. IEEE.

Laskar, Zakaria, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. 2017. "Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Network." *arXiv Preprint arXiv:1707.09733*.