Deep SfM

Report (Work in Progress)

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.

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.

Review of SfM

Common SfM pipelines can be categorized in two:

1. Feature based approaches -

- Extraction of keypoints and features across images and matching them to obtain correspondences
- Estimating pairwise relative camera motion from the correspondences
- Recovering 3D structure from the camera motion and features

2. Direct approaches that directly solve for structure and camera motion from tha images.

Feature based methods are very sensitive to outliers and noise in feature matching while direct approaches are computationally very expensive.

An excellent survey of the SfM is presented in Özyeşil et al. (2017) .

Key challenges

Feature matching

Obtaining dense correspondences by matching features is usually the first step in a standard SfM pipeline. However, this usually fails when there are images with low textures, complex geometry or occlusions. There has been some success in alleviating this with deep learning (Han et al. 2015)

Globally inconsistent solution

Most approaches rely on solving smaller tractable sub-problems like pair-wise frame pose and structure estimation and then optionally optimize them globally. Whilst obtaining local consistency gives good initial estimates for the global optimization, it is still an intractable problem which is solved to certain threshold.

Lack of semantic priors

Given two images seeing the same structure from two different viewpoints, humans can easily estimate their relative camera pose and a general 3D structure of the scene. This is due to the prior knowledge we possess from our experiences seeing the world. Deep learning based approaches have shown tremendous progress in being able to learn a similar prior despite the need to have lots of data, finely constructed architectures and hand-tuned training regimens.

Goals

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 test sequence of images from a scene different than the training set, the model should be able to predict a plausible explanation of the 3D structure and poses.
- Interpretable 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 allow supervision from scene-geometry based signals (ex, re-projection error).
- Scalable: The approach must be independent 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.

Why deep learning?

Deep neural networks can potentially leverage visual priors and help combat deficiencies in existing pipelines. To that end, our model should be able to learn how 2D images relate to the scene structure and changes in viewpoint affect the projected image. Being able to learn such strong priors should allow it to estimate viewpoint and 3D structure from 2D images alone.

Problem formulation

Given a sequence of images $\{I_1, I_2, ..., I_n\}$ as the input, extract the following output:

- 3D scene structure defined by $P = \{p_1, p_2, ..., p_m\}$ where $p_i \in \mathcal{R}^3 \times \mathcal{R}^3 \times \mathbb{R}^3$ is a 3D point-sample storing a position, normal and color values, respectively.
- Camera poses for each input image $\{c_1, c_2, ..., c_n\}$ where $c_i \in \mathcal{R}^3 \times \mathcal{R}^4$ is the 3D camera position and 3D camera rotation in quaternion representation of image I_i.

However, the above formulation can be further simplified by noting that each 3D point can be mapped to one or more 2D pixels in the input images (we focus only on the geometry of the scene and not the illumination effects in the scene, ie, we ignore the rendering aspects of the problem). Let us explore different routes to simplify the redundancy in the pointcloud output

Mathematical perspective

Is the mapping injective from the input space to output sp ace?

Given an input, is the output a distinct one? The above simple formulation does not account for unique output (SfM ambiguity problem). Here's an example: Let us assume we a 3D point cloud P with m 3D points. We also have two camera poses $C = \{C_1, C_2\}$. Let I_1, I_2 be the images projected on the camera at C_1 and C_2 respectively. So if we were to solve for the camera poses and point cloud with I_1, I_2 as our input to the model, P, c is a valid solution. However, this solution is not unique. If we translate all the points by a vector x to get a new point cloud $P' = \{p'_i = p_i + x \forall p_i \in P\}$ and similarly, translate all the camera poses in C by x, it would still be a valid solution. In fact, it can be shown that if we transform the scene by any non-singular 4×4 matrix T, then we can apply the same transformation to the camera poses to obtain the same camera-space projections (the images I)[Refer Appendix for proof]

Let us now update the problem to remove this SfM ambiguity.

Engineering perspective

Appendix

Proof of SfM ambiguity

Given a group of m 3D points $P = \{p_1, p_2, ..., p_m\}$ viewed by n cameras with camera poses $C = \{C_1, C_2, ..., C_n\}$, then the homogeneous projection coordinate of the j^{th} point onto the k^{th} camera is :

$$\begin{aligned} i_k^j &= C_k^{-1} p_j && (1)\\ &= C_k^{-1} T^{-1} T p_j && (2)\\ &= (T C_k)^{-1} (T p_j) && (3) \end{aligned}$$

Thus any non-singular matrix T can be applied to the scene and the camera poses to get the same projections. Note: camera pose matrix is the inverse of the extrinsic matrix.

References

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.

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.

Han, Xufeng, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. 2015. "Matchnet: Unifying Feature and Metric Learning for Patch-Based Matching." In *Computer Vision and Pattern Recognition (Cvpr), 2015 Ieee Conference on*, 3279–86. IEEE.

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*.

Özyeşil, Onur, Vladislav Voroninski, Ronen Basri, and Amit Singer. 2017. "A Survey of Structure from Motion*." *Acta Numerica* 26. Cambridge University Press:305–64.