3D Space-Time Correspondence as a Contrastive Random Walk

Linyi Jin Changyuan Qiu Zhuowen Shen Yuliang Zhu
University of Michigan
{jinlinyi,peterqiu,mickshen,yuliangz}@umich.edu

1. Introduction

Humans have a remarkable ability to understand dynamic scenes from videos including estimating the 3D structure as well as finding correspondence at different time stamps. For example, when opening a fridge, one can track the door of the fridge and estimate the state and shape of the fridge. Many prior works [23, 14, 11] infer the 3D structure of a static scene from videos. Other lines of works [22, 29, 22, 7] solve the "what went where" [31] problem of dynamic scenes. However, estimating the 3D structure from videos of *dynamic* scenes is still a challenging problem. The classical method Structure from Motion (SfM) is limited to only static scenes [23, 24, 15]. In addition, video datasets with annotations for temporal visual correspondences are scarcely available and hard to create, making supervision a bottleneck for object tracking [10, 32].

In this work, we propose to estimate the 3D structure of dynamic scenes by combining a 3D detection system with a self-supervised visual temporal correspondence learning system. We reconstruct plane structures from video frames and track the reconstructed planes in dynamic scenes. The core methodology of this work is highly inspired by [7] from Jabri *et al*. We incorporate 3D planar surfaces information obtained from PlaneRCNN [12] of each frame into the framework of [7], which offers important geometric constraint for patch affinity prediction in dynamic scenes.

2. Related Works

Structure from Motion is a well-studied field that reconstructs 3D structures based on 2D images. However, prior works [25, 4, 30, 26, 23, 24, 15] are limited to static scenes and cannot handle the occlusions caused by dynamic objects. Visual SLAM systems [30, 20, 21] capture consistent room-scale and surface-based maps using RGB-D images or videos. They find point or plane correspondences across frames and use geometry to predict the camera location of each frame and a static 3D pointcloud. Aside from using sensor depthmaps, they also assume certain parts of the environment remain unchanged. Our approach, on the other hand, aims to use RGB frames only and work on dynamic scenes.

Recently, many works use deep learning methods to reconstruct 3D structure from images. There have been works to reconstruct normals [5, 28], voxels [2, 6], and depth [5, 19] from 2D images. However, single view 3D cannot guarantee consistency across video frames and cannot handle dynamic scenes. There are also efforts to estimate consistent depth from video frames using temporal correspondences [9, 16, 27, 35] or geometric correspondences [14]. Our approach differs from these works by estimating planes instead of depth, which is a higher level representation.

Our approach builds most heavily on works aiming to produce a planar reconstruction [13, 33, 12, 34, 1, 8]. In particular, we build on PlaneRCNN [12], which detects planes along with their geometric properties and instance masks from a single RGB image. Despite using surrounding frames to improve the detection quality, it only addresses the detection based on a single frame, and does not touch on the application when the temporal correspondence of the targeted image sequence needs to be learned.

[29, 7] aims to learn useful visual representations for visual correspondence from raw videos without any supervision. Our work builds on [7] which enforces strong affinity between image patches by learning embedding vectors to guide a random walk across a palindrome of frames. However, the original work suffers from very strict formulation to prevent the network from finding shortcuts during learning. Our approach adds plane normal aside from the embedding vectors to construct the affinity between nodes of adjacent frames. Our hope is that adding extra constraints can help the system to find robust planar correspondence across frames.

3. Approach

We will use PlaneRCNN [12] to acquire the plane segmentation and the corresponding plane normal vectors. Then we will combine the normal vectors with the 2D patches in [7] to improve robustness of tracking.

3.1. Planar Patch and Normal Detection

The planar patches are obtained from the refinement network and the plane parameters are obtained from the plane detection module.

3.2. Normal Assisted Random Walk

We incorporate the plane parameters (normal, offset) into the 2D frame patches in the random walk process of [7]. Only planar patches that has high intersection-over-union (IoU) with the randomly sampled patch will be assigned the plane parameters.

We assume that the camera displacement between the two frames is not drastic, so that corresponding patches have similar normals. Then our new stochastic matrix of affinities would be:

$$A' = A + \lambda F(\text{normal difference}), \tag{1}$$

where A is the affinity in [7] and function F is some geometric function calculating the similarity of the normal of the two adjacent frames when normal exists for both patches. We will leave the rest of our network the same as in [7].

4. Evaluation

We will train the PlaneRCNN [12] on the ScanNet [3]. We will evaluate our 3D plane detection on raw ScanNet videos and report VOI, RI, SC and AP for planes follow [12]. We will evaluate object tracking on DAVIS 2017 [18] and report the mean (m) and recall (r) of standard boundary alignment (F) and region similarity (J) metrics, which are detailed in [17].

References

[1] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *CVPR*, 2020. 1

[2] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 1

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2

[4] Ashwin P Dani, Nicholas R Fischer, and Warren E Dixon. Single camera structure and motion. In *TACON*, 2011. 1

[5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 1

[6] R. Girdhar, D.F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 1

[7] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. In *NeurIPS*, 2020. 1, 2

[8] Ziyu Jiang, Buyu Liu, Samuel Schulter, Zhangyang Wang, and Manmohan Chandraker. Peek-a-boo: Occlusion reasoning in indoor scenes with plane representations. In *CVPR*, 2020. 1

[9] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. In *TPAMI*, 2014. 1

[10] Xueting Li, Sifei Liu, Shalini de Mello, Xiaolong Wang, Jan Kautz, and Ming Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *NeurIPS*, 2019. 1

[11] Chen-Hsuan Lin, Oliver Wang, Bryan C Russell, Eli Shechtman, Vladimir G Kim, Matthew Fisher, and Simon Lucey. Photometric mesh optimization for video-aligned 3d object reconstruction. In *CVPR*, 2019. 1

[12] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *CVPR*, 2019. 1, 2

[13] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *CVPR*, 2018. 1

[14] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. In *SIGGRAPH*, 2020. 1

[15] Kemal E Ozden, Konrad Schindler, and Luc Van Gool. Multibody structure-from-motion in practice. In *TPAMI*, 2010. 1

[16] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. In *RA-L*, 2020. 1

[17] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2

[18] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. 2

[19] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In *TPAMI*, 2020. 1

[20] Carolina Raposo and Joao P Barreto. πMatch: Monocular vslam and piecewise planar reconstruction using fast plane correspondences. In *ECCV*, 2016. 1

[21] Carolina Raposo, Miguel Lourenço, Michel Antunes, and João Pedro Barreto. Plane-based odometry using an rgb-d camera. In *BMVC*, 2013. 1

[22] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1

[23] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[24] Gideon Schwarz et al. Estimating the dimension of a model. In *Annals of statistics*, 1978. 1

[25] Sudipta N Sinha, Drew Steedly, and Richard Szeliski. A multi-stage linear approach to structure from motion. In *ECCV*, 2010. 1

[26] Mohamed Tamaazousti, Vincent Gay-Bellile, Sylvie Naudet Collette, Steve Bourgeois, and Michel Dhome. Nonlinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment. In *CVPR*, 2011. 1

[27] Rui Wang, Stephen M Pizer, and Jan-Michael Frahm. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In *CVPR*, 2019. 1

[28] Xiaolong Wang, David F. Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *CVPR*, 2015. 1

[29] Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 1

[30] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. In *RSS*, 2015. 1

[31] Josh Wills, Sameer Agarwal, and Serge J. Belongie. What went where. In *CVPR*, 2013. 1

[32] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 1

[33] Fengting Yang and Zihan Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *ECCV*, 2018. 1

[34] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piece-wise planar 3d reconstruction via associative embedding. In *CVPR*, 2019. 1

[35] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *ICCV*, 2019. 1