Review for Unsupervised Learning of Depth and Ego-Motion from Video

Irem Arpag

March 7, 2021

The article deals with creating a model similar to humans' capability of inferring ego-motion and 3D structures of a scene in real-time, by locating obstacles and reacting them readily, the field I which almost no improvement has performed within years in computer vision, especially on the areas where occlusion or non-rigidity are existing. One approach explains this noticeable distinction of human as the result of past visual experiences that are gained moving around and watching very large number of scenes, in that way improve a coherent model that helps us to sense simply when we see a new scene. To be inspired by this understanding, the authors design an end-to-end conv. neural network, consisting of two independent models, the first is Depth CNN for depth prediction and the second Pose CNN for pose estimation, and formulate the model to map directly from input pixels to an estimate of ego-motion while training it with sequence of images that are unlabeled video clips.

The significance of the model is under the idea training, the system with an unsupervised learning method in which images only come from monocular videos, the authors argue that it is one of the first in the computer vision world.

The aim of the research is to provide not a perfect or clear, but a consistent or intermediate explanation of the visual world in terms of depth prediction and pose estimation by encouraging the pipeline on vast amount of unstructured video sequences, no labeling even no camera motion information, that is the cause of very blurred images generally.

Comparing the first qualitative results of the experiment by means of methodology and system, it is regarded that the said unsupervised model is worthy of comparison with supervised models and pose estimation performs advantageously in accordance with established SLAM systems. The system that has three prediction modules composed of depth, pose and explainability estimation networks also evaluate with prior frameworks through using mainly the KITTI dataset and additionally City-Spaces and Make 3D datasets for improving purposes. The most challenging point here is that depth estimation evaluation of the unsupervised model is made with supervised models that also use calibrated stereo images. For single-view depth estimation, the researchers first pre-trained the system on the larger City-Spaces dataset, the result of which have some structural errors such as holes on car surface, and then they fine-tune on KITTI that improve the results slightly, but still better than supervised models esp. preserving depth boundaries and thin structures. For depth estimation, the quantitative performance of the model is good enough to compare with the current supervised methods except the work of Godard and his colleagues that also creates an idea for future challenge. Pose estimation results that are tested with KITTI odometry split are compared with two variants of SLAM system, ORB-SLAM (short) and ORB-SLAM (full) that uses more data performs better than the said model. The visualization performance of the explainability prediction shows that the CNN has learned to identify dynamic objects such as pedestrians or objects disappear from the frame as unexplainable with success, but it cannot be valid for thin structures that the network has problem to identify, similarly the ablation study on explainability modeling cannot offer prosperous results.

As future work, the researchers think to extend current aggregation of the model to different fields such as object detection and semantic segmentation. The approach for unsupervised learning of depth and ego-motion by using only monocular videos inspired many other researchers such as with their similar setting, Vijayanarasimhan et al [1] tries to learn motion of the objects in the scene and in another work,

Mahjourian et al [2] presents differentiable 3D loss functions establishing consistency between the geometry of adjacent frames for improving depth and ego-motion estimation. The paper is a detailed, easy to understand, comprehensive one, diversified with vast sample images comparisons.

REFERENCES

[1] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar and K. Fragkiadaki. Sfm-net:Learning of Structure and motion from video. Arxiv:2017

[2] Mahjourian R, Wicke M, Angelova A., Unsupervised Learning of Ego-motion from Monocular Video using 3D Geometric Constraints, IEEE 2018.