Supplementary Material for Joint Semantic Segmentation and 3D Reconstruction from Monocular Video

Abhijit Kundu, Yin Li, Frank Daellert, Fuxin Li and James M. Rehg

Georgia Institute of Technology, Atlanta, USA
http://www.cc.gatech.edu/~akundu7/projects/JointSegRec

Additional video results, data and related code are avialable at the project website[1]. Here we provide additional discussion on the experiments (§ 1) and some details of the datasets (§ 2) used in the experiments.

1 Experiments Discussion

Output of our system is a 3D *semantic+occupancy* map. However due to lack of ground truth in that form, we need to evaluate using indirect approaches. To evaluate the segmentation accuracy, we evaluated it with standard 2D semantic segmentation methods for which human annotated ground truth exists. The 2D segmentation is obtained by back-projecting our 3D map to the camera images. However these kind of evaluation negatively harms our scores for the following reasons:

- Dynamic objects (cars, pedestrians) present in these annotated images negatively hurts our results for all categories as these objects rightly does not occur in our static 3D map. Upon 2D backprojection we see the categories behind these moving objects which is considered wrong according to 2D segmentation benchmarks.
- Discretization effects due to voxel size of our 3D map, especially those close to camera trajectory negatively harms per pixel 2D segmentation accuracy. Some of these effects can be seen on the video. In other words, due to projective nature of the camera, even a error in one voxel close to camera path causes multiple number of pixels to be miss-classified.
- Ground truth annotations are only available for a sparse set of images only in video datasets like [1,4]. So quantitative improvements are not as drastic as one would expect from a fully temporally consistent segmentation. This fact has been noted by other authors [6]. Note that in Table1 in main paper, for Leuven we have no segmentation scores for *Pedestrian* but we have consistency scores. This is because some voxels are labeled as pedestrian, indeed temporally consistent, but none are visible from the annotated images.

However in-spite of all the above issues, we out-perform the state of art in per pixel 2D segmentation accuracy. Some of the above issues can be properly

[1] http://www.cc.gatech.edu/~akundu7/projects/JointSegRec

addressed by a dataset which has Ground Truth for semantic segmentation in 3D space. However this remains a future work.

2 Dataset Details

Details of the image sequences used for experiments and runtime are listed in Table. 1. We implemented the system in C++. Current unoptimized runtime for the system is around 20 minutes on standard desktop machine for 800 images long sequence involving about 20 million voxels.

Dataset	Image Resolution	Trajectory Length	Avg. Runtime
CAMVID [1]	960x720	Every 800 images (approx. 0.4km)	20 mins
LEUVEN [2,4]	316x256	1174 images (approx. 0.6km)	19mins

Table 1. Details of the datasets and system runtime.

Our semantic segmentation evaulation (in main paper) on CAMVID is on a subset (seq05VD) of the whole CAMVID dataset[1]. Semantic segmentation scores for KITTI [3] are based on odometry evaluation sequence 05. We used the latest code by Ladicky et al.[5,4] and results of [6] and [7] for Camvid seq05VD were provided by their authors.

References

1. Brostow, G., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. PRL 30(2), 88–97 (2009)
2. Cornelis, N., Leibe, B., Cornelis, K., Van Gool, L.: 3d urban scene modeling integrating recognition and reconstruction. IJCV 78(2-3), 121–141 (2008)
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
4. Ladicky, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., Torr, P.H.: Joint optimisation for object class segmentation and dense stereo reconstruction. In: BMVC (2010)
5. Ladicky, L., Russell, C., Kohli, P., Torr, P.: Associative hierarchical crfs for object class image segmentation. In: ICCV (2009)
6. Miksik, O., Munoz, D., Bagnell, J.A., Hebert, M.: Efficient temporal consistency for streaming video scene analysis. In: ICRA (2013)
7. Tighe, J., Lazebnik, S.: Superparsing: Scalable nonparametric image parsing with superpixels. International Journal of Computer Vision (2012)