Supplementary Material - CarFusion: Combining Point Tracking and Part Detection for Dynamic 3D Reconstruction of Vehicles*

N Dinesh Reddy Minh Vo Srinivasa G. Narasimhan
Carnegie Mellon University
{dnarapur,mpvo,srinivas}@cs.cmu.edu

In this document, we provide additional materials to supplement our main submission. The method applies to any rigidly moving objects. The formulation can be further generalized to piece-wise rigid or articulated objects. We do not correspond the unstructured points across multiple wide-baseline videos but only track them within a single video. Only structured points are corresponded across videos.

Structured and Unstructured Points

The structured points are 14 car keypoints, obtained by training the Stacked hourglass CNN architecture [2] on the annotated dataset of Kitti dataset and applying this model to our videos. Multi-View Bootstrapping is the process of finetuning the Stacked hourglass CNN model on our videos using the reprojected locations of the 14 3D car keypoints to the all cameras (views) as self-supervised labels. Tab. 3 in the paper shows the improvement in the CNN model when applied to our videos and the resulting improvement in 3D reconstruction of the structured points. The unstructured points are detected using Harris corner detection algorithm and tracked using a extended version of lucas-kanade tracker.

Comparison with CAD/Active Shape Model Fitting

Figure 1 shows a comparison between a CAD dictionary fitting algorithm of ref [1] applied to a single image and our approach. CAD fitting approaches are sensitive to errors in 2D keypoint localization, especially in the presence of occlusions. In unconstrained settings as ours, CAD fitting orientation error is approximately 22 degrees (while our method shows only 1 degree error). Further, since current methods were trained on a small range of CAD models, they cannot generalize to arbitrary vehicles in the wild.

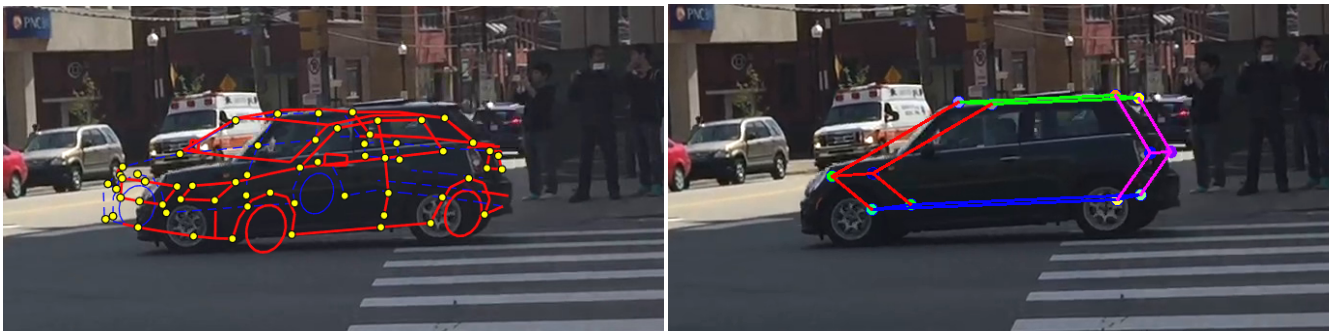

Figure 1: [Left] CAD based reconstruction [1] shows 22^o angular error with respect to the motion of the vehicle. [Right] In comparison, our result shows 1^o angular error. The CAD basis based methods fail due to inaccurate keypoint predictions, especially when only one side of the vehicle is visible.

*http://www.cs.cmu.edu/ ILIM/projects/IM/CarFusion/

Comparison with single video methods

Figure 2 shows a comparison between our method and a traditional single-video based SFM reconstruction of structured points on the intersection dataset. Single video based reconstruction fails as can been seen from the reprojection of structured points onto a different viewpoint (Where single video methods have 35 pixels RMS reprojection error versus ours of 2.5 pixels).

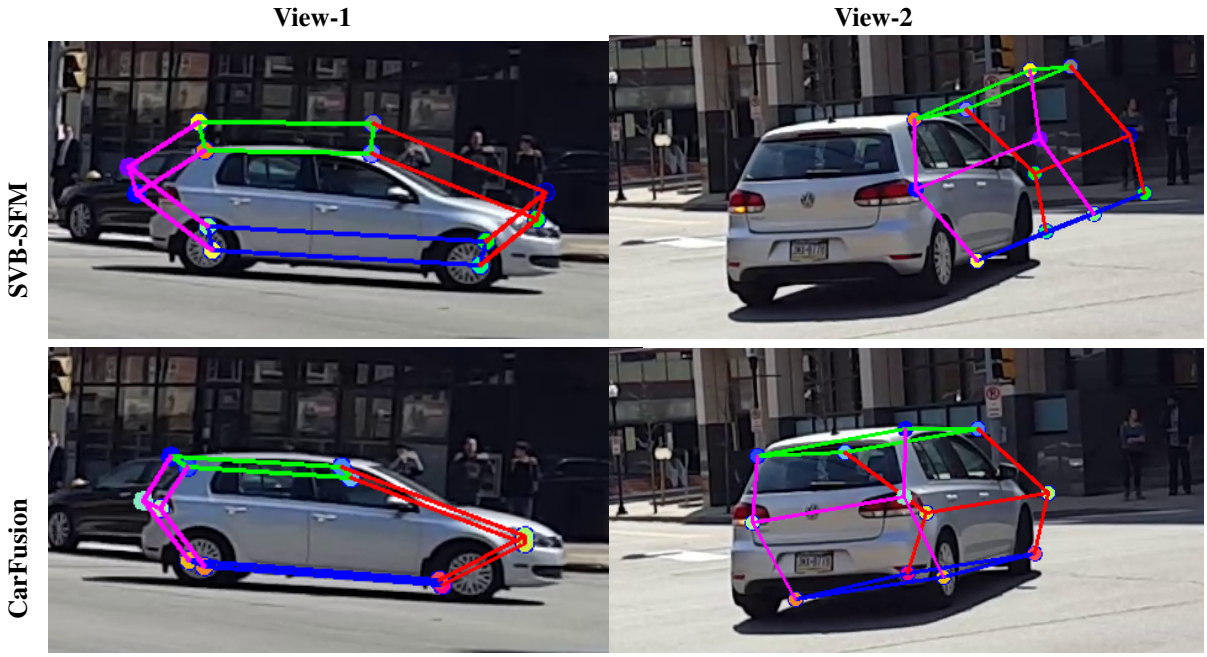

Figure 2: [Top-Left] Single-video based SFM(SVB-SFM) reconstruction of structured points. [Top-Right] The reprojection of the reconstruction onto a different view. [Bottom] Projection of reconstructed car using CarFusion to the above corresponding views.

Evaluation

We evaluate our method extensively on nearly 210,000 frames (3 minutes of 21 videos, each with 10000 frames), which is about 10 times the data (6 minute single video) evaluated by state-of-the-art SLAM pipelines (e.g., LSD-SLAM of Engel et al.). We show RMS reprojection errors in Tables 1 and 2 (in paper). Here we assume the structured points are accurately predicted by the network and evaluate the reconstruction. Tab. 3 compares the predicted structure points against ground truth labels. In lieu of actual 3D ground truth locations, RMS reprojection error in completely different views has been the most widely used evaluation metric in SFM and SLAM pipelines. We also show 3D reconstructions of many cars registered to a satellite photograph of the intersection.

References

[1] Y.-L. Lin, V. I. Morariu, W. Hsu, and L. S. Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *European Conference on Computer Vision*, pages 466–480. Springer, 2014.

[2] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.