# Supplementary Material for the Paper "Monocular Multiview Object Tracking with 3D Aspect Parts"

Yu Xiang[1,2⋆], Changkyu Song[2⋆], Roozbeh Mottaghi[1], and Silvio Savarese[1]

[1] Computer Science Department, Stanford University
{yuxiang, roozbeh}@cs.stanford.edu, ssilvio@stanford.edu
[2] Department of EECS, University of Michigan at Ann Arbor
changkyu@umich.edu

## 1 Dataset Description

Table 1 shows the length and the viewpoint change in terms of azimuth, elevation and distance of the 9 videos in our new car tracking dataset and the 11 sequences from the KITTI dataset [3]. Race1 to Race6 are video clips of racing cars recorded by a moving camera, which contain severe appearance change due to viewpoint transition, occlusion by tire smoke, and high speed motion blur. SUV1 and SUV2 are test driving videos which show all possible viewpoints of the SUVs and contain severe occlusions due to snow and dust respectively. Sedan is a camera record of learning car-drift, where the car turns around a track repeatedly. KITTI01 to KITTI11 are sequences captured by a camera mounted on a moving car. There can be multiple cars in each sequence, but we specify one car to track, which changes its pose in the sequence. Besides, in some sequences, the target is occluded by other objects temporarily. We can see from the table that the viewpoints of the cars change dramatically, which poses a big challenge to tracking.

## 2 Annotation Process

To provide ground truth annotations for viewpoints and 3D aspect parts, we use the pose annotation tool proposed in [4]. We associate each car with a 3D CAD model downloaded from Google 3D Warehouse [1]. Fig. 1 shows the four CAD models we used for the YouTube and the KITTI sequences, where we collected two sedans and two SUVs. Then we identify 12 anchor points in the 3D CAD models for car, which are displayed as red circles in Fig. 1. In 2D, we annotate the locations of these anchor points in the video frames. Finally, using the 2D-3D correspondences of the anchor points, we can compute accurate viewpoints and 3D aspect part locations for the targets by minimizing the re-projection error between the projected anchor points and the annotated anchor points.

⋆ indicates equal contribution.

| Video | Length | Azimuth | Elevation | Distance |
|---|---|---|---|---|
| Race1 | 153 | 217.80 / 1.42 | 53.32 / 0.35 | 30.93 / 0.20 |
| Race2 | 138 | 234.61 / 1.70 | 57.99 / 0.42 | 53.26 / 0.39 |
| Race3 | 111 | 196.65 / 1.77 | 56.34 / 0.51 | 8.00 / 0.07 |
| Race4 | 69 | 114.73 / 1.66 | 28.44 / 0.41 | 13.63 / 0.20 |
| Race5 | 146 | 269.47 / 1.85 | 94.35 / 0.65 | 30.82 / 0.21 |
| Race6 | 231 | 460.85 / 2.00 | 124.19 / 0.54 | 26.35 / 0.11 |
| SUV1 | 219 | 209.41 / 0.96 | 17.86 / 0.08 | 5.71 / 0.03 |
| SUV2 | 851 | 818.25 / 0.96 | 76.24 / 0.09 | 62.49 / 0.07 |
| Sedan | 700 | 606.80 / 0.87 | 73.31 / 0.10 | 72.55 / 0.10 |
| KITTI01 | 28 | 66.03 / 2.36 | 21.19 / 0.76 | 36.85 / 1.32 |
| KITTI02 | 27 | 92.83 / 3.44 | 32.32 / 1.20 | 39.28 / 1.45 |
| KITTI03 | 60 | 259.64 / 4.33 | 69.40 / 1.16 | 73.06 / 1.22 |
| KITTI04 | 29 | 80.24 / 2.77 | 24.15 / 0.83 | 36.25 / 1.25 |
| KITTI05 | 43 | 120.30 / 2.80 | 22.10 / 0.51 | 41.85 / 0.97 |
| KITTI06 | 49 | 71.80 / 1.47 | 39.04 / 0.80 | 26.82 / 0.55 |
| KITTI07 | 32 | 78.98 / 2.47 | 36.35 / 1.14 | 22.18 / 0.69 |
| KITTI08 | 27 | 86.37 / 3.20 | 10.67 / 0.40 | 13.55 / 0.50 |
| KITTI09 | 30 | 123.15 / 4.11 | 66.19 / 2.21 | 45.68 / 1.52 |
| KITTI10 | 66 | 134.95 / 2.04 | 33.37 / 0.51 | 27.36 / 0.41 |
| KITTI11 | 31 | 129.91 / 4.19 | 19.82 / 0.64 | 16.61 / 0.54 |

**Table 1.** The statistics of the 9 videos in our new car dataset and the 11 sequences from the KITTI dataset [3]. Length is the number of frames in the video. For the last three columns, the first number in each cell indicates the accumulated azimuth/evelation/distance change in the video, while the second number indicates the average azimuth/evelation/distance change per frame, where the unit of azimuth and elevation is degree and unit one in distance corresponds to the size of the car in 3D.

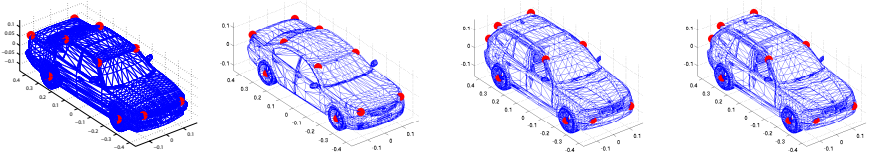

**Fig. 1.** CAD models used in the annotation process.

## 3 Result Video

Please see the project page [2] for the tracking results in our experiments.

## References

1. Google 3d warehouse. http://sketchup.google.com/3dwarehouse
2. http://cvgl.stanford.edu/projects/multiview_tracking
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
4. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: WACV (2014)