Recognizing Car Fluents from Video
Supplementary Material

Bo Li[1,*], Tianfu Wu[2], Caiming Xiong[3,*] and Song-Chun Zhu[2]
[1]Beijing Lab of Intelligent Information Technology, Beijing Institute of Technology
[2]Department of Statistics, University of California, Los Angeles [3]Metamind Inc.
boli86@bit.edu.cn, {tfwu, sczhu}@stat.ucla.edu, cmxiong@metamind.io

1. Main Challenges on Car-Fluent Dataset

Our Car-Fluent dataset includes 647 video clips, containing basically 10 types of semantic parts and 16 types of car part fluents with diverse camera viewpoints and occlusion conditions. Fig. 3 shows the whole scene context of these videos. The videos are collected from various sources (youtube, movies, VIRAT [3], etc.), and captured by both static cameras and moving cameras. As can be seen from Fig. 3, there are both high and low resolution parts, which pose great challenges on part localization and status estimation.

On this dataset, the **semantic parts** are:

- "hood"
- "left-front door"
- "left-back door"
- "right-front door"
- "right-back door"
- "trunk"
- "left head light"
- "right head light"
- "left tail light"
- "right-tail light".

The **car fluents** are:

- "open/close left-front door"
- "open/close left-back door"
- "open/close right-front door"
- "open/close right-back door"
- "open/close hood"
- "open/close trunk"
- "change left/right lane"
- "turn left/right".

Main challenges on this dataset are all related to car parts and fluents, including[1]:

1. the large geometry and appearance variation of cars introduced by part status change;
2. low resolution of car parts;
3. diverse occlusion introduced by people;
4. the variation of fluent execution rate;
5. diverse viewpoints.

The first three figures in Fig. 1 show the relative positions of different semantic parts (which are color coded), the variance of part size, and the variance of part aspect ratio on this dataset. Here, part position and part size are normalized by the size of the whole car, i.e., car body. As can be seen, each semantic part has disordered distribution and large variances of part size, or part aspect ratio. This is because an opened part is very different from a closed one in terms of size, aspect ratio, and relative positions w.r.t car body.

For each semantic car part, we also plot the heat map of its distribution, which can be seen in Fig. 2. We can see the parts are distributed diversely because of the status change, and there are several "peaks" reflect the principal positions of the "open" parts and "close" parts.

*This work was done when Bo Li was a visiting student and Caiming Xiong was a Postdoc at UCLA.

[1]For car lights, main challenges are the low resolution and ambiguous appearance, as they don't have the geometry change, but have periodically intensity change, we omit these parts in the following analysis.

On the last of Fig. 1, we show the number of opening frames compared to the ones of closing frames, which can be viewed as the temporal variance of fluent videos on proposed dataset. The big variance is caused by the diverse execution rate. For instance, the speed of opening left-front door depends on different people, and even for the same person, the speed is not always equal. Other challenges include intra-class variations of fluent change on different car types (e.g., opening the trunk of a jeep is very different from the same fluent on a sedan) and background clutters.

Based on the experimental results in our paper, we can see these challenges pose a hard problem to current vision models. Since these videos are captured from real scenarios, we believe it is suitable for fluent recognition and part status estimation in the wild, and hope this could draw more attentions in our community.

2. Car-Fluent Videos

We show the challenging videos on Car-Fluent dataset, please check them in the directory - "video-demos". Since there are many cars that don't have fluent change in the original video, we ask the annotators to only annotate cars that have fluent change. To simplify the annotation, annotators may choose not to annotate the moving cars in the parking lot scenario.

For each video, we show the "close/turn-off" parts by solid rectangle, the "open/turn-on" parts by dashed rectangle, and the process of car fluents change by dotted rectangle. When there are fluents change, we also show the text of specific fluents name on the right of each video.

For the car lights video, we only show the whole car bounding boxes for better visualization.

3. Current Performance

As can be also seen Table 1 in our paper, the overall performance of fluent recognition is still low. In our experiments, we find our model detect some wrong part status, or just missed when parts are heavily occluded or too small, and thus get wrong transition spatial-temporal features. For STIP and iDT, we find they missed capturing part motions on some videos, and many features are often extracted on people. This probably because car parts are small, ambiguous with background, and often occluded by people. Overall, more informative features and modelling strategies are needed to cope with these cases in the future.

4. About Part Localization Baselines

We also try to compare our model with Yang and Ramanan's mixtures-of-parts model [4], and its CNN extension [2], but find it's hard to compare with them. There are mainly 3 reasons: first, the parts in their model are equally-sized. However, in real life, different semantic car parts have different sizes and aspect ratios, especially when the parts are opened. We found it's not easy to extend their framework to model diverse sizes and aspect ratios of parts; second, it's not very easy to model a fully-connected skeleton model as human pose for car. For example, when facing a car in the frontal view, it is very hard to speculate the bounding box of the tail lights. Thus it's hard to design a fully-connected skeleton model for car. Third, they need fine scale landmark annotations for input while there is no such annotation on Car-Fluents Dataset, besides, parts in their model are not "functional" (e.g, may not be open).

In literature, [1] also proposed a fully-collected part model which is related to our work. But their code is not released, and we found it's not easy to re-implement their model.

5. More Part-Localization and Part-Status Estimation Results

In addition to the results of part localization and status estimation in our paper, we show more results here.

Fig. 4 shows more successful examples of our ST-AOG. Fig. 5 shows more failure ones. For better visualization, we only show the cropped detection results of cars. We can see our model can localize parts and estimate their statuses fairly well with different viewpoints and car types. The failure cases are mainly due to occlusion, car view misdectection, low part resolution, or background clutters. We will cope with these problems in the future work. From these detection results, we can also see the position of a part can change a lot when its status change.

References

[1] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014.

[3] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011.

[4] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

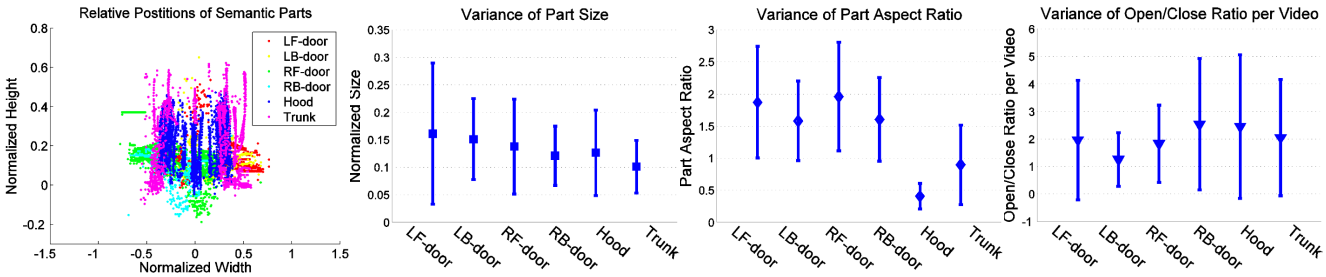

Figure 1. Some statistics of semantic parts on Car-Fluent dataset. In the first figure, different parts are coded with different colors. In the second and third figures, we show the variances of part size and aspect ratios. These two statistics reflect the geometry variations of car parts on Car-Fluent dataset. The last figure shows the "open/close" ratios of car parts, which reflects the diverse execution rates of car fluents.

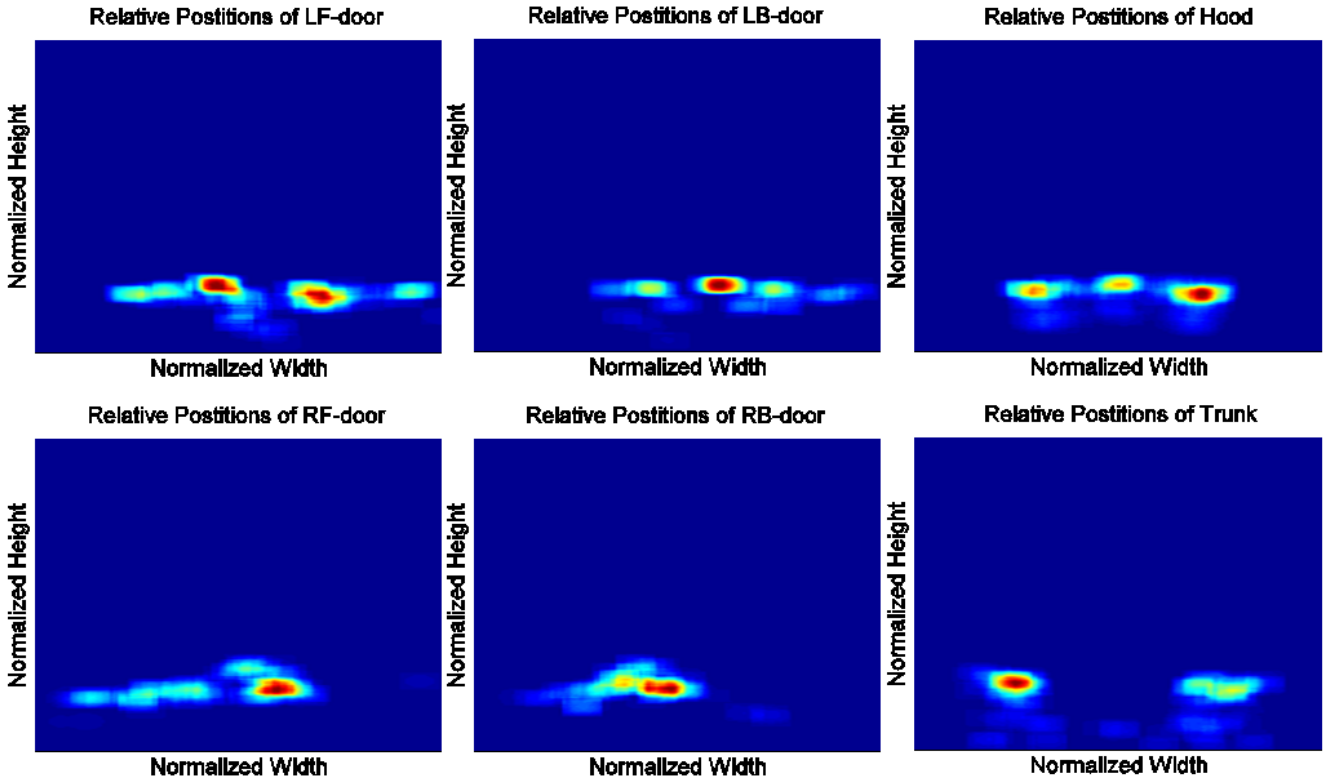

Figure 2. Semantic part distributions on Car-Fluent dataset. We align each semantic part with the whole car position, and normalize the part size with the whole car bounding box. We can see several "peaks" in the distribution of each semantic part. These "peaks" reflect the principal positions of "open" parts and "close" parts.

Figure 3. Sample images of our Car-Fluent Dataset, from which we can see the diverse scene context of car fluents.

Figure 4. Successful detection examples of ST-AOG on Car-Fluent dataset. For good visualization, we crop the cars from the original detection images, and show the car lights separately. Different semantic parts are shown by rectangles in different colors (solid for "close", or "turn-off" status, and dashed for "open", or "turn-on" status). As can be seen our ST-AOG is fair in localizing these parts and estimating the corresponding statuses.

Figure 5. Failure detection examples of ST-AOG. For good visualization, we crop the cars from the original detection images, and show the car lights separately. Different semantic parts are shown by rectangles in different colors (solid for "close", or "turn-off" status, and dashed for "open", or "turn-on" status). The black lines are used to illustrate variations of the relative positions between whole car and car parts with different statuses. The failure cases are mainly due to the occlusions introduced by people, the mis-detected viewpoints, low part resolutions, or there are background clutters.