Supplementary Material: Joint Inference of Groups, Events and Human Roles in Aerial Videos

Tianmin Shu¹, Dan Xie¹, Brandon Rothrock², Sinisa Todorovic³ and Song-Chun Zhu¹ Center for Vision, Cognition, Learning and Art, University of California, Los Angeles

{stm512, xiedan}@g.ucla.edu sczhu@stat.ucla.edu

²Jet Propulsion Laboratory, California Institute of Technology

brandon.rothrock@jpl.nasa.gov

³School of Electrical Engineering and Computer Science, Oregon State University

sinisa@onid.orst.edu

1. Dataset



Figure 1: Image of our hex-rotor in the air with a GoPro camera.

We assembled a new low-cost hex-rotor with a GoPro camera shown in Fig. 1, which is able to eliminate the high frequency vibration of the camera and hold in air autonomously through a GPS and a barometer. It can also fly $20 \sim 90m$ above the ground and stays 5 minutes in air. We use this hex-rotor to take a set of videos with some plots at a park where the terrain is interesting: hiking routes, parking lots, camping sites, picnic areas with shelters, restrooms, tables, trash bins and BBQ ovens. By detecting/tracking humans and objects in the videos, we can recognize events such as BBQ, queuing, exchanging objects, loading/unloading, etc.

We have collected some events with scripts involving the interactions between humans and objects. Fig. 2 shows two frames captured from the original videos. The original videos are pre-processed, including camera calibration and frame registration. After pre-processing, there are totally 27 videos in the dataset, the length of which ranges from 2 minutes to 5 minutes. We annotate the hierarchical semantic information of objects, roles, events and groups in the



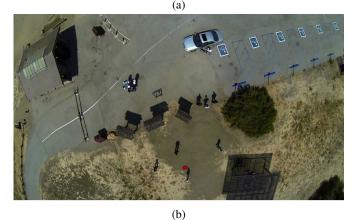


Figure 2: Two frames of the original aerial videos from two different sites.

videos. Tab. 1 is a summary of events, roles, objects and the number of instances in our dataset.

ID	Event	Objects	Roles	# of instances
1	Exchange Box	Box, Car	Deliverer, Receiver	11
2	Play Frisbee	Frisbee	Player	13
3	Info Consult	Desk, Info Booth	Consultant, Visitor	11
4	Pick Up	Car	Driver, Passenger	9
5	Queue for Vending Machine	Vending Machine	Queuing Person	9
6	Group Tour	N/A	Guide, Tourist	6
7	Throw Trash	Trash Bin	Thrower	12
8	Sit on Table	Table & Seat	Customer	10
9	Pinic	Blanket	Picnic Person	4
10	Serve Table	Table & Seat	Waiter, Customer	6
11	Sell BBQ	BBQ Oven	Chef, Buyer	6

Table 1: Summary of the dataset.

Object type	Precision	Recall
Buildings	100.0%	95.65%
Cars	10.38%	30.64%
Small static objects	16.16%	53.33%

Table 2: Object detection accuracy. Small static objects include *Table & Seat*, *Info Booth*, *Desk*, *BBQ Oven*, and *Trash Bin*. We have accurate building detection results while the detection accuracy for small static objects and cars is not very ideal, which affects the recognition of events that involve these objects.

Precision	Recall	Rate of broken tracklets
31.36%	39.05%	4.44

Table 3: Tracking accuracy. Tracking generates a large number of false alarms and fails when humans/objects have little motions, which can be indicated by the low precision and recall. The rate of broken tracklets shows that ground truth trajectories are broken into multiple fragments in the tracking results. The poor tracking results greatly increase the difficulty of our inference.

2. Detection and Tracking Evaluation

The accuracy of the detection of buildings, cars and small static objects is shown in Tab. 2.

Tab. 3 shows the accuracy of tracking.