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

Tianmin Shu[1], Dan Xie[1], Brandon Rothrock[2], Sinisa Todorovic[3] and Song-Chun Zhu[1]
[1]Center for Vision, Cognition, Learning and Art, University of California, Los Angeles [2]Jet Propulsion Laboratory, California Institute of Technology [3]School of Electrical Engineering and Computer Science, Oregon State University

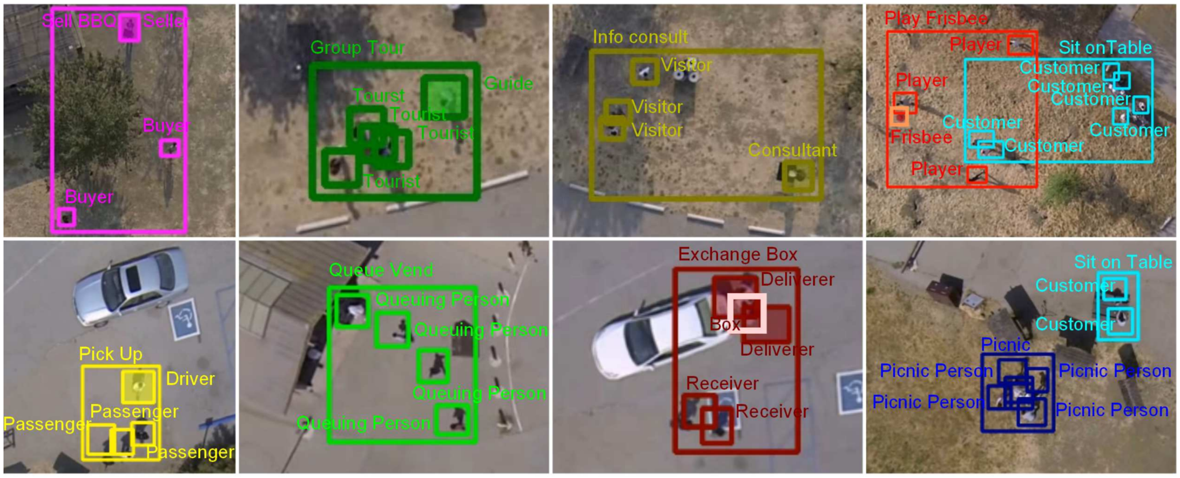

Figure 1: Visualization of results including groups (large bounding boxes), events (text) and human roles (small bounding boxes with text). In events with more than one role, we use the shaded bounding box to represent the second role; small portable objects are labeled with lighter color. From event and human role recognition, we can group people even when they are far from each other (e.g.,*Play Frisbee* and *Sell BBQ*).

Video surveillance of large spatial areas using unmanned aerial vehicles (UAVs) is becoming increasingly important in a wide range of civil, military and homeland security applications. Yet, there is scant prior work on aerial video analysis [1, 2, 4], which for the most part is focused on tracking people and vehicles (with few exceptions [3]) in relatively sanitized settings.

Towards advancing aerial video understanding, this paper presents a new problem of parsing low-resolution aerial videos of large spatial areas, in terms of grouping and assigning roles to people and objects engaged in events, and recognizing these events. Given an aerial video, our objectives include: 1) grouping people based on their events, 2) recognizing events present in each group, and 3) recognizing roles of people involved in these events.

In this paper, domain knowledge is formalized as spatiotemporal AND-OR graph (ST-AOG). Its nodes represent the following four sets of concepts: events $\Delta_{\mathrm{E}} = \{E_i\}$; sub-events $\Delta_L = \{L_a\}$; human roles $\Delta_{\mathrm{R}} = \{R_j\}$; small objects that people interact with $\Delta_{\mathrm{O}} = \{O_j\}$; and large objects and scene surfaces $\Delta_{\mathrm{S}} = \{S_j\}$. A particular pattern of foreground trajectories observed in a given time interval gives rise to a sub-event, and a particular sequence of sub-events defines an event. ST-AOG has special types of nodes. An AND node, $\wedge$, encodes a temporal sequence of latent sub-events required to occur in the video so as to enable the event occurrence (e.g., in order to *Exchange Box*, the *Deliverer*s first need to approach the *Receiver*s, give the *Box* to the *Receiver*s, and then leave). For a given event, an OR node, $\vee$, serves to encode alternative space-time patterns of distinct sub-events.

A temporal segment of foreground trajectories corresponds to a sub-event, which is represented as the *latent* spatiotemporal template of n-ary spatiotemporal relations among foreground trajectories within a time interval. In each sub-event, foreground trajectories form characteristic space-time patterns that can be robustly extracted from training videos through unsupervised clustering.

A parse graph is an instance of ST-AOG, explaining the event, sequence of sub-events, and human role and object label assignment. The solution of our video parsing is a set of parse graphs, $W = \{pg_i\}$, where every pg_i explains a subset of foreground trajectories, $G_i \subset G$, as

$$pg_i = \{e_i, \tau_i = [t_{i,0}, t_{i,T}], \{L(\tau_{i,u})\}, \{\boldsymbol{r}_{i,j}\}\}, \tag{1}$$

where $e_i \in \Delta_{\mathrm{E}}$ is the recognized event conducted by G_i; $\tau_i = [t_{i,0}, t_{i,T}]$ is the temporal extent of e_i in the video starting from frame $t_{i,0}$ and ending at frame $t_{i,T}$; $\{L(\tau_{i,u})\}$ are the templates (i.e., latent sub-events) assigned to non-overlapping, consecutive time intervals $\tau_{i,u} \subset \tau_i$, such that $|\tau_i| = \sum_u |\tau_{i,u}|$; and $\boldsymbol{r}_{i,j}$ is the human role or object class assignment to jth trajectory $\Gamma_{i,j}$ of G_i.

Our objective is to infer W that maximizes the log-posterior $\log p(W|G) \propto -\mathcal{E}(W|G)$, given all foreground trajectories G extracted from the video.

Given an aerial video, we first build a video panorama and extract foreground trajectories G. Then, the goal of inference is to: (1) partition G into disjoint groups of trajectories $\{G_i\}$ and assign label event $e_i \in \Delta_{\mathrm{E}}$ to every G_i; (2) assign human roles and object labels $\boldsymbol{r}_{i,j}$ to trajectories $\Gamma_{i,j}$ within each group G_i; and 3) assign latent sub-event templates $L(\tau_{i,u}) \in \Delta_L$ to temporal segments $\tau_{i,u}$ of foreground trajectories within every G_i. For steps (1) and (2) we use two distinct MCMC processes. Given groups G_i, event labels e_i and role assignment $r_{i,j}$ proposed in (1) and (2), step (3) uses dynamic programming for efficient estimation of sub-events $L(\tau)$ and their temporal extents τ. Steps (1)–(3) are iterated until convergence.

We have prepared and released a new aerial video dataset [1]. The dataset contains 27 videos, 86 minutes, 60 fps, resolution of 1920 × 1080, with about 15 actors in each video. All video frames are registered onto a reference plane of the video panorama and provided with detailed annotation of groups, events, and human roles.

We evaluate our approach on both annotated bounding boxes (grouping: 95.48%, event: 96.38%, role: 89.94%) and real tracking results (grouping: 49.47%, event: 32.84%, role: 18.92%). Our results demonstrate that we successfully address above inference tasks under challenging conditions.

[1] Yumi Iwashita, M.S. Ryoo, Thomas J. Fuchs, and Curtis Padgett. Recognizing humans in motion: Trajectory-based aerial video analysis. In *BMVC*, 2013.

[2] M. Keck, L. Galup, and C. Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In *WACV*, 2013.

[3] Omar Oreifej, Ramin Mehran, and Mubarak Shah. Human identity recognition in aerial images. In *CVPR*, 2010.

[4] Jan Prokaj and Medioni Gerard. Persistent tracking for wide area aerial surveillance. In *CVPR*, 2014.

[1]Dataset can be download from http://www.stat.ucla.edu/~tianmin.shu/AerialVideo/AerialVideo.html