Unsupervised Learning of Activities in Video using Scene Context

Sangmin Oh  
Kitware Inc.  
sangmin.oh@kitware.com

Anthony Hoogs  
Kitware Inc.  
anthony.hoogs@kitware.com

Abstract

Unsupervised learning of semantic activities from video collected over time is an important problem for visual surveillance and video scene understanding. Our goal is to cluster tracks into semantically interpretable activity models that are independent of scene locations; most previous work in video scene understanding is focused on learning location-specific normalcy models. Location-independent models can be used to detect instances of the same activity anywhere in the scene, or even across multiple scenes. Our insight for this unsupervised activity learning problem is to incorporate scene context to characterize the behavior of every track. By scene context, we mean local scene structures, such as building entrances, parking spots and roads, that moving objects frequently interact with. Each track is attributed with large number of potentially useful features that capture the relationships and interactions with a set of existing scene context elements. Once feature vectors are obtained, tracks are grouped in this feature space using state-of-the-art clustering techniques, without considering scene location. Experiments are conducted on webcam video of a complex scene, with many interacting objects and very noisy tracks resulting from low frame rates and poor image quality. Our results demonstrate that location-independent and semantically interpretable groupings can be successfully obtained using unsupervised clustering methods, and that the models are superior to standard location-dependent clustering.

1 Introduction

Unsupervised learning of semantic activities from video collected over time is an important problem for visual surveillance and video scene understanding. An example scene of OceanCity, NJ, USA is shown in Figure 1, captured by a surveillance webcam camera for more than an hour. It can be seen that moving objects, mostly people and vehicles, have been automatically tracked by a visual tracker where the collected trajectories are overlaid on the scene. The learning of semantic activities in this work refers to the problem of grouping tracks into semantically similar categories such as people walking on sidewalks, people crossing roads, vehicles leaving parking lots, and vehicles driving straight, in an unsupervised manner. The semantic grouping problem is qualitatively different from previous work on trajectory analysis [3, 8, 5, 7, 9] in that every discovered group of activities in our work aim to be semantically coherent, regardless of event locations. For example, trajectories that belong to people crossing roads will result in one single cluster using our approach while multiple clusters will be formed per event locations based on previous work [3, 8, 5, 7, 9]. In addition to the superior interpretability, the advantage of obtaining semantically coherent and location-independent clusters is that resulting models are potentially transferable to different scenes. For example, assuming that we have a location-independent model of people crossing roads, it can be used to retrieve similar activity instances from trajectories collected at different sites, effectively alleviating the problem of learning site-specific models and resulting in a more scalable solution.

Without scene location, however, trajectory information can be too impoverished for useful semantic behavior characterization. Our solution is to incorporate scene context to characterize the behavior of every track. By scene context, we mean local scene structures such as building entrances, parking spots and roads, which moving objects often interact with. Semantic activities are usually defined by the relations and interactions between moving objects and scene elements. For example, the activity of people walking into roads can be characterized by attributes such as move_on_sidewalk, move_towards_road, move_on_road, and move_at_human_speed. Accordingly, each track is encoded with large number of potentially useful features which capture its relationships and interactions w.r.t. a set of scene elements.
Once track features are computed, we grouped tracks using four different clustering techniques: K-means, mean shift [1], spectral clustering [10], and affinity propagation [2]. From our experimental results, we found that spectral clustering [10] and affinity propagation [2] yield superior groupings which allow semantically plausible interpretations and substantial separability with minimal parameter tuning efforts. Activities such as people walking on sidewalks and vehicle parkings are identified, and tracks belonging to each activity are grouped together regardless of their event locations.

Previous work has considered low-level scene elements such as track sources and sinks [8, 7, 9] and normal traffic flows computed from tracks. However, higher-level scene content is becoming increasingly accessible through commercial services, e.g., Google Earth and Microsoft Virtual Earth. Furthermore, they can be automatically inferred from data. For example, Swears and Hoogs [6] presented supervised learning methods to detect specific scene elements such as building entrances and parking spots.

The remainder of the paper is organized as follows. Sec. 2 describes data, scene contexts, and associated features. Then, Sec. 3 describes clustering techniques and demonstrates experimental results.

2 Data, Scene Contexts, Features

The video data used in this work is shown in Fig. 1 where it was recorded from a webcam located in Ocean City, NJ, USA. The frame rate is almost always 1Hz or worse, the pixel-level noise is very high, and there is often horizontally-banded noise. There is significant per-frame compression, such that artifacts are clearly visible and degrade detection performance.

We compute tracks automatically by detecting moving objects and performing global, multi-object tracking, following a previous work [4]. Note that the unsupervised semantic activity learning presented in this work does not depend on the tracker, and in principle any video detection and tracking system could be used.

Once tracks are computed, the next level of representation is the characterization of tracks by single-track features, relations between tracks and the scene, and local, short-term actions involving tracks. This level of information should capture all salient aspects of localized object behavior for trajectory analysis. The manner and timing of interactions between a moving object and other static scene elements can indicate significantly different types or behaviors. We have focused on exploring the best strategies to exploit the available functional scene contexts rather than learning and identifying them from video sources. In this work, five different types of scene contexts which moving objects frequently interact with are manually identified and marked: roads, sidewalks, parking spots, building entrances, and trash cans. All five color-coded scene contexts on image plane are shown in Fig. 2.

To characterize the spatial relations and actions between tracks and existing scene contexts both on image and world coordinate systems, our approach captures the changes in distance within every trajectory to the nearest scene context locations. For most surveillance datasets, the geometry between image and real-world coordinates is available or computable, e.g., homography. Figure 3 shows the pre-computed context distance maps in both image and world coordinates for trash cans shown in Fig. 2. Using these distance maps in conjunction with the tracking results, we can compute action/relation binary features per track efficiently.

In detail, total 31 binary features which can be broadly categorized into three categories are computed.
for every track, as enlisted in Table 1: track-level, contextual, and composite features. The third type of composite features are designed based on heuristics to roughly categorize tracks to be human or vehicle, based on factors such as speed, location, and bounding box sizes. In detail, the threshold speed of 6.5 m/s was used as the boundary between $F_3$ and $F_4$. Additionally, the heuristic features $F_{30}$ and $F_{31}$ are defined as follows:

$$F_{30} = F_1 \& F_4 \& \{SO(S)|E||P)||RC\}$$

$$F_{31} = F_2 \& F_3 \& \{SO(R)||P)||RC\}$$

where $RC = \{MA(S)||P)&MT(R) \}|| {MT(S)||P)&MA(R)\}$. Here, $RC$, $SO$, $MA$, $MT$, $S$, $E$, $P$, and $R$ denote road_crossing, stay_on, move_away, move_toward, sidewalk, entrance, parking spot, and road respectively. Note that the binary features are not mutually exclusive; both composite features can be true, indicating that a track maybe both human and a vehicle, embracing the uncertainty and mitigate more strict decision to other computational modules. The track features described above are computed for entire trajectories, yielding a one vector per track.

## 3 Track Clustering

The computed set of vectors are clustered using four different methods which include both standards and state-of-the-arts: $K$-means, mean-shift [1], spectral clustering [10], and affinity propagation [2]. In particular, affinity propagation [2] has shown to learn clusters better separated than standard $K$-means. Each cluster represents a group of tracks that share a common set of features. While $K$-means explicitly requires to specify the number of target clusters in advance, the other three methods have alternative sensitivity parameters which determine aggressiveness in track grouping. Experiments were repeated with different parameter settings on total 3252 tracks collected from a video set which totals 86 minutes.

In our experiments, spectral clustering [10] and affinity propagation [2] produce semantically interpretable clusters with minimal parameter tuning efforts. Although numerous parameter tunings were tried for the other two methods, we could not obtain semantically plausible results. In general, parameters have been tuned to yield between 10 and 25 clusters. In one result, affinity propagation produced 11 clusters, and spectral clustering produced 22 clusters with fairly high semantic interpretability. Figure 4(a-d) shows tracks within four sample clusters among total eleven classes identified by affinity propagation. It can be observed that each class delivers interpretable semantic behaviors: vehicles passing through, vehicles parking, people walking on sidewalks, and people crossing roads, all independent of event locations. In our knowledge, such highly interpretable location-independent semantic grouping results have not been reported using unsupervised approaches. It shows that the use of scene contexts and proposed set of relation/action features are useful for unsupervised semantic activity learning.

As a comparison, a trajectory analysis method similar to [7] has been conducted on the same dataset. First, the entire 2D world-coordinates were divided into grids with fixed size of 2 by 2 meters. Then, every track was encoded into a normalized bag-of-words histogram where each word corresponds to a particular grid. Then, affinity propagation is used with sensitivity parameters tuned to yield 15 clusters. Two sample clusters obtained by this comparative method are shown in Figure 4(e-f). It can be observed that the clusters are mostly domi-

<table>
<thead>
<tr>
<th>ID</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Tracking bounding box size indicates person?</td>
</tr>
<tr>
<td>F2</td>
<td>Tracking bounding box size indicates vehicle?</td>
</tr>
<tr>
<td>F3</td>
<td>Fast moving (within normal vehicle speed)?</td>
</tr>
<tr>
<td>F4</td>
<td>Slow moving (within normal human speed)?</td>
</tr>
<tr>
<td>F5-9</td>
<td>Stay on scene contexts within world?</td>
</tr>
<tr>
<td>F10-14</td>
<td>Move nearby scene contexts within world?</td>
</tr>
<tr>
<td>F15-19</td>
<td>Move nearby scene contexts within image?</td>
</tr>
<tr>
<td>F20-24</td>
<td>Move away from scene contexts within world?</td>
</tr>
<tr>
<td>F25-29</td>
<td>Move toward scene contexts within world?</td>
</tr>
<tr>
<td>F30</td>
<td>Possibly a human?</td>
</tr>
<tr>
<td>F31</td>
<td>Possibly a vehicle?</td>
</tr>
</tbody>
</table>

Table 1. List of track features in three categories: (1-4) track-level, (5-29) contextual, and (30-31) composite.
nated by spatial distribution, and fail to deliver substantial semantic groupings. For example, the cluster shown in Fig. 4(e) is a large mixture of tracks that occur on or nearby roads, which indiscriminately include multiple activities: vehicles driving straight, vehicles parking, and people crossing roads. Furthermore, Fig. 4(f) only shows pedestrian tracks on the left sidewalk, but fails to group with similar tracks on the right sidewalk.

4 Conclusion

We have studied the impact of using scene context features for unsupervised semantic activity learning. The experimental results are promising and demonstrate that location-independent semantic groupings can be successfully obtained, aided by the use of scene context. In future work we plan to use automatically-generated scene context, and to study the generalization of learned activities across different scenes.

References