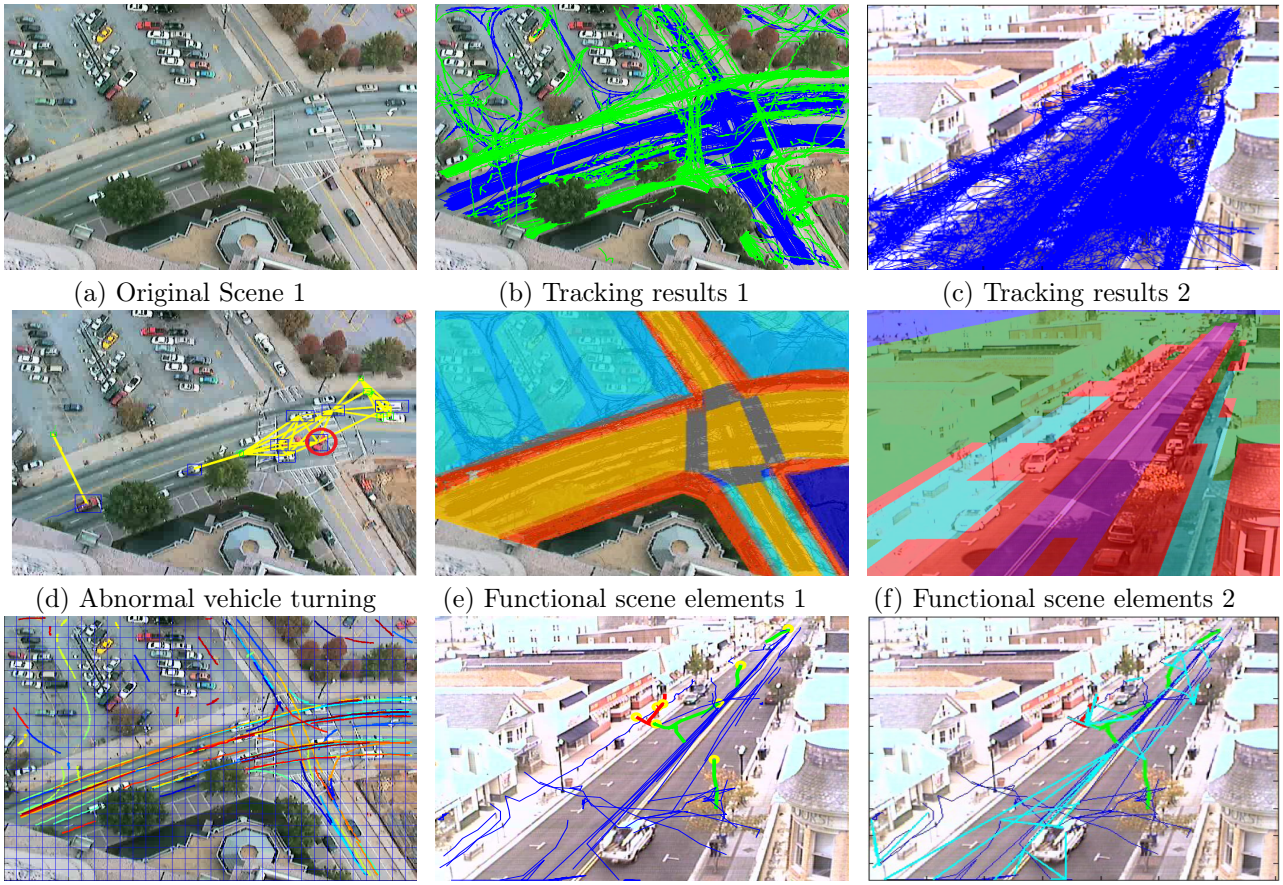
Anomaly detection from videos under sparse data and partial observations*

Sangmin Oh, Anthony Hoogs
{sangmin.oh, anthony.hoogs}@kitware.com
Kitware Inc. 28 Corporate Dr., Clifton Park, NY

Anomaly detection from videos is an important computer vision problem where the goal is to identify rare activity examples which significantly deviate from normal behaviors observed in scenes. For example, Fig 1 shows a scene (a) and tracking results (b), where vehicles (blue) and people (green) are tracked over an extended period of time. An example anomaly is shown in Fig. 1(d) where a vehicle (red circle) is turning at a traffic intersection confronting oncoming traffic.



(a) Original Scene 1 (b) Tracking results 1 (c) Tracking results 2

(d) Abnormal vehicle turning (e) Functional scene elements 1 (f) Functional scene elements 2

(g) Activity words of vehicles (h) Fragmented tracks (i) Linked tracklets

Figure 1: (a) Scene 1. (b) Tracking results in scene 1 with vehicles (blue) and people (green). (c) Tracking results from scene 2. (d) Abnormal instance of a vehicle turning towards incoming traffic at an intersection. (e) Manually marked functional scene elements from scene 1. (f) Automatically grouped scene elements from scene 2. (g) Activity words of vehicles obtained by grouping track BOWs. (h) Examples of fragmented tracking results. (i) Automatically linked tracklets by learned linking function.

*Topic: visual processing and pattern recognition. Preference: poster.

In this work, we address the problem of anomaly detection under sparse data and partial observations. In video surveillance, we often have sufficient training data to learn basic normalcy models such as standard motion patterns. For example, the dominant motion patterns of scenes 1 and 2 in Fig. 1 are evident in the computed tracks in Fig 1 (b) and (c), which were computed on tens of minutes and a few hours of video respectively. However, there is often insufficient data to learn less common patterns that still occur regularly and are therefore not anomalies. If we desire to learn higher-order models, such as those including vehicle type or appearance, then considerably more data is required; the curse of dimensionality and training set imbalances become serious issues.

Anomaly detection is further exacerbated by frequent partial observations due to tracking errors. Tracking algorithms can miss objects entirely, lose an object after tracking it for a while, switch between objects, or virtually any combination of these. An example of a delivery truck and crew from scene 2 is shown in Fig. 1 (h) where the starting locations of fragmented tracklets are marked by yellow circles along with the tracks belonging to crew (red), truck (green), and other simultaneous tracks (blue).

Many state-of-the-art video scene understanding approaches [7, 9, 3] adopt the paradigm of scene-specific analysis where they learn location-specific normalcy models based on video exclusively from the scene, which are used to detect anomalies based on likelihoods against the learned models. Some of them use normalcy modeling over a large hypothesis space, which is reduced through mid-level, quantized representations such as bag-of-words [9]. Activity words learned from vehicle tracks from scene 1 are shown in Fig. 1(g). A practical limitation of these approaches is that a large amount of data is required per scene, and the learned models are not generalizable to novel scenes; i.e. there is no mechanism to transfer knowledge from one scene to another. Furthermore, these approaches either assume complete tracks, and fail with major tracking errors; or use low-level motion directly, and cannot detect long-range anomalies.

A promising approach to alleviate partial observation problems is to link tracks using normalcy models. Recent learning-to-link approaches demonstrated that track linking can be improved by a linking function [4] that can be learned per functional mover type [6]. An example linking result [6] for the delivery truck example in Fig. 1(h) is shown in Fig. 1(i). However, it is not clear whether the previously learned track linkers can benefit the new track linker in a novel scene at all.

In this work, we propose to compensate for sparse training data and tracking errors by incorporating functional scene elements [8, 5, 6] into anomaly detection and track linking. Functional scene elements (FSEs) are local scene regions with different functionalities such as straight roads, intersections, sidewalks, crosswalks, and parking spots. Moving objects tend to behave analogously within identical FSEs. Examples of FSEs from two scenes are shown in Fig. 1(e) and (f), marked by different colors. FSEs provide us with the ability to match functionally similar regions between two scenes, allowing us to transfer normalcy from one to another. Accordingly, an anomaly detection system can perform more effectively for a novel scene, even with limited amount of scene-specific data. A track linker can be benefited as well because successful tracks from corresponding FSE regions within other scenes can provide useful cues to link tracklets in a novel scene. The previous video scene understanding [7, 9, 3] and learn-to-link [4, 6] approaches adopted batch-mode learning framework which require large amount of scene-specific data and do not naturally extend to incremental learning. Accordingly, we also investigate the use of on-demand local learning approaches [1, 2, 10] for both anomaly detection and track linking problems. Our preliminary results will be shown in complex video surveillance scenes, where anomalies would be detected on previously-unseen scenes, or scenes with only a few minutes of video.

References

[1] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11(1-5):11–73, 1997.

[2] James Hays and Alexei Efros. Scene completion using millions of photographs. In *SIGGRAPH*, 2007.

[3] Daniel Küttel, Michael D. Breitenstein, Luc J. Van Gool, and Vittorio Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010.

[4] Y. Li, C. Huang, and R. Nevatia. Learning to associate : Hybridboosted multi-target tracker for crowded scene. In *CVPR*, 2009.

[5] Turek M., Hoogs A., and Collins R. Unsupervised learning of functional categories in video scenes. In *ECCV*, 2010.

[6] Sangmin Oh, Anthony Hoogs, Matthew W. Turek, and Roderic Collins. Content-based retrieval of functional objects in video using scene context. In *ECCV (1)*, 2010.

[7] Imran Saleemi, Lance Hartung, and Mubarak Shah. Scene understanding by statistical modeling of motion patterns. In *CVPR*, 2010.

[8] Eran Swears and Anthony Hoogs. Functional scene element recognition for video scene analysis. In *IEEE Workshop on Motion and Video Computing*, 2009.

[9] Xiaogang Wang, Keng Teck Ma, Gee Wah Ng, and Eric Grimson. Trajectory analysis and semantic region modeling using a nonparametric Bayesian model. In *CVPR*, 2008.

[10] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *ECCV*, 2010.