Content-based Retrieval of Functional Objects in Video using Scene Context

Sang Min Oh and Anthony Hoogs
Kitware Inc.
28 Corporate Drive, Clifton Park, New York 12065
[sangmin.oh|anthony.hoogs]@kitware.com

Functional object recognition in video is an emerging problem in visual surveillance and video scene understanding. By functional objects, we mean objects with a specific purpose such as a postman or delivery truck, which are defined more by their actions and behaviors than by appearance. Examples of delivery and trash trucks are shown in **Figure 1**. The two major challenges that arise from using videos are: (1) learning functional models that capture location-independent behavioral semantics, and (2) track fragmentation due to imperfect visual tracking. In this work, we present an approach for content-based learning and recognition that effectively addresses these issues.

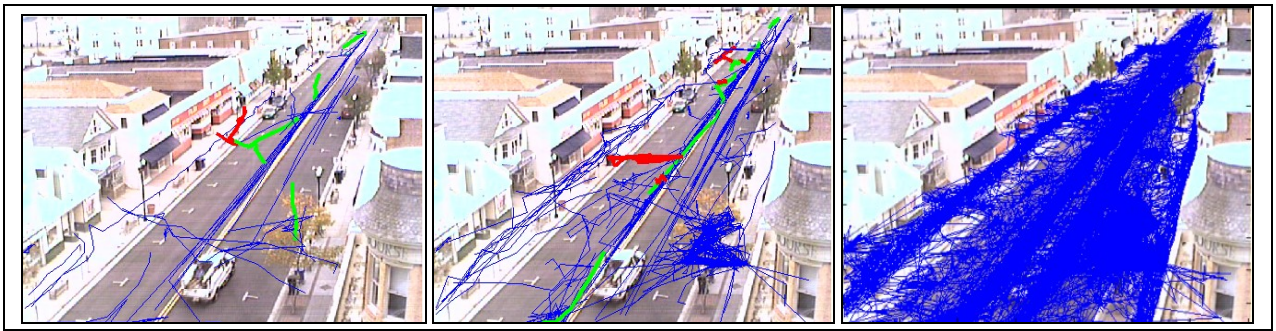

Figure 1. Examples of Functional objects: (Left) delivery truck, and (Middle) trash truck, among (Right) all tracks. Tracks belonging to human and vehicle movers are shown in red and green respectively for delivery and trash truck examples where fragmentation is clearly visible. Blue tracks are other tracks that occurred concurrently.

We have formulated a two-level representation to model functional object behavior over time in the presence of track fragmentation. At the lower level, all the collected tracks are clustered based on features relating them to scene elements, resulting in models corresponding to different categories of low-level behaviors such as "walking on sidewalk" and "crossing road". At the higher level, composite functional object models are learned in a supervised fashion, using the low-level elementary functional models as building blocks. Each track is quantized based on learned clusters and higher-level models are learned from these quantized data, abstracting away low-level information. Full positive examples are given in the form of manually linked tracks exhibiting each function. In terms of modeling regimes, we have investigated three approaches: (1) unigrams, (2) bigrams, and (3) HMMs.

Our solution for location-independent semantic low-level behavior learning is to incorporate *scene context* to characterize tracks. By scene context, we mean local scene regions with different functionalities such as doorways, parking spots and roads, which moving objects often interact with (see **Figure 2**). Every track is encoded with Boolean features which capture its relations and actions w.r.t. existing scene contexts, e.g., 'MoveTo' or 'AwayFrom'. 39 features are computed and used to cluster tracks into elementary functional behaviors. Two example clusters (among 11 total clusters) are shown in **Figure 2**, with reasonable high-level semantic interpretations. Our semantic grouping results are qualitatively different from previous work on trajectory analysis where tracks are primarily grouped based on their location information [1][2]. Multiple clustering approaches have been explored: K-means, mean-shift, affinity propagation [4], and spectral clustering [3]. We have found affinity propagation to produce semantically interpretable results with minimal parameter tuning efforts.

Figure 2. Left – manual scene context: road (green), parking spots (light blue), sidewalks (yellow), doorways (red), and trash cans (dark blue). Learned elementary behavior clusters: (middle) parking and (right) walking on sidewalk.

Higher level functional models are learned from positive examples of linked tracks exhibiting each behavior. Histograms of quantized low-level behaviors are used for unigram and bigram models, where the latter encodes sequential pairs of behaviors. We also explored HMMs of behavioral words.

Recognition in the presence of track fragmentation is addressed by a track linking classifier using standard Adaboost.M1 [5] trained on manually-labeled pairs of tracks from the same object as positive examples, and random pairs as negative examples. Instead of individual track features, vectors encoding the similarity between pairs of track features are used. In our test, the learned link classifier delivered 99% recall and false-alarm rate of 0.9%. **Figure 3** shows an example result with both correct recalls and false alarms.

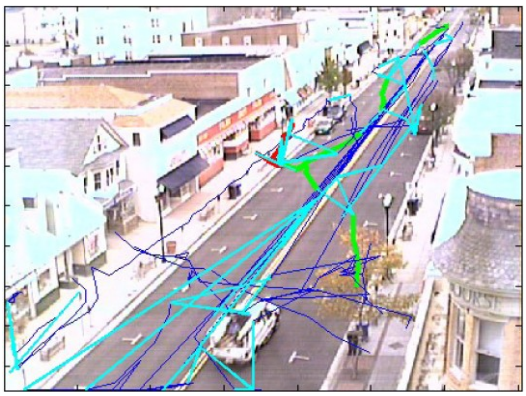

Figure 3. Detected links overlaid (cyan) on the example of a delivery truck shown in **Figure 1**.

Sequences of tracks with higher link probabilities are formed into functional behavior hypotheses. Each linked hypothesis is evaluated against the full functional behavior models, using Bhattacharyya distances for unigrams and bigrams, and data-likelihood for HMMs. In our experiments, there were 20-2500 hypotheses for ~10 minutes of video, depending on functional class. We used uncontrolled webcam video of a retail street scene, with significant tracking difficulties from low frame rates (1-2Hz), low resolution and severe noise. We considered four functional categories: delivery person, delivery truck, road cleaning vehicle, and trash truck. Experiments were conducted in leave-one-out fashion with 7 to 15 exemplars per category. In addition, experiments were repeated using automatically learned scene context, which is less accurate but more desirable. The ROC curve for the delivery truck class and average percentile of true examples among all generated linked hypotheses are shown in **Figure 4**. An important finding is that simpler models of unigrams and bigrams consistently generate more accurate recognition results than more complex HMMs; this is likely due to the limited number of training examples.

Given the difficulty of the functional recognition problem in real-world settings, these initial results are promising results. We plan to apply the approach to diverse scenes and larger number of functional object categories.

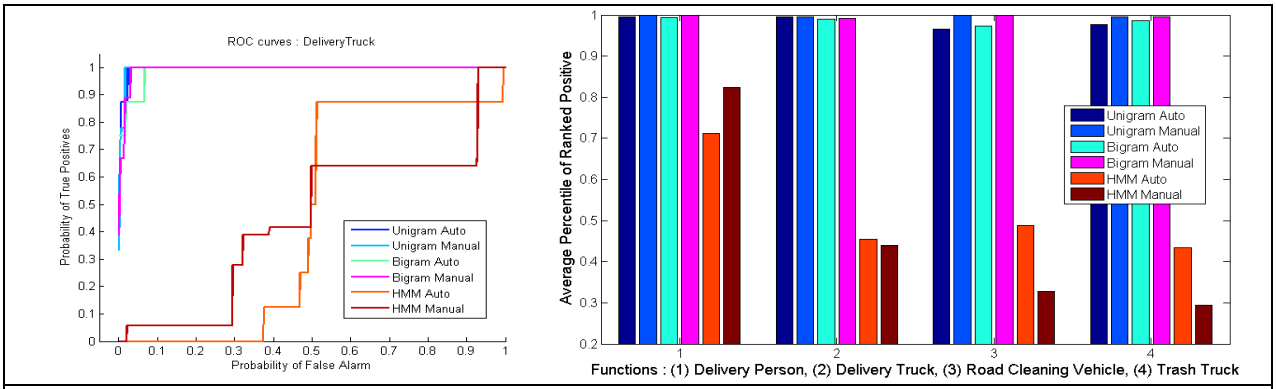

Figure 4. (Left) ROC curves for delivery truck class. 'Auto' refers to the use of less-accurate learned scene context clusters, while 'Manual' refers to manual labels. (Right) Average ranking ratios of true samples for all categories.

[1] E. Swears, A. Hoogs, and A. G. A. Perera**.** "Learning Motion Patterns in Surveillance Video using HMM Clustering." *Proceedings of the IEEE Workshop on Motion and Video Computing,* 2008.

[2] C. Stauffer, W.E.L. Grimson, "Learning Patterns of Activity Using Real-Time Tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22 (8), 747--757, 2000.

[3] L. Zelnik-Manor and P. Perona. "Self-Tuning Spectral clustering." In Proc. Neural Information Processing (NIPS) 18, 2006.

[4] B. J. Frey and D. Dueck. "Clustering by passing messages between data points." Science, Vol. 315, No 5814, pages 972-976, Feb. 2007.

[5] Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of Computer and System Sciences, 55(1):119-139, 1997.