Unsupervised Learning of Functional Scene Elements in Video Scenes

Matt Turek, Anthony Hoogs, and Roderic Collins
Kitware, Inc.
28 Corporate Drive, Clifton Park, NY 12065
[matt.turek|anthony.hoogs|roddy.collins]@kitware.com

We present a novel form of video scene analysis where functional scene element categories are learned. Many functional scene elements, such as roads, parking areas, sidewalks, and entrances, can be segmented and categorized based on the behaviors of moving objects in and around them, while distinguishing them based on appearance is very difficult. Existing work in video scene modeling has largely focused on segmenting dominant motion patterns [1, 2, 4] and significant regions such as track sources and sinks [3, 4], given observed trajectories and detection algorithms for each scene element type. Our work differs in that we do not attempt to segment the various motion patterns in a scene from each other, or to develop detection algorithms specific to any scene element category. Instead, our method consists of: 1) developing a common feature-space representation for all behavioral scene element categories; 2) learning behavioral category models using the representation, in an unsupervised fashion, independent of scene location; 3) segmenting video scenes into the functional categories.

We view the problem from the perspective of categorical object recognition. Our approach identifies regions (functional scene elements) with similar behaviors in the same scene and/or across scenes, by clustering histograms based on a trajectory-level, behavioral codebook. A cluster of such objects corresponds to a functional category that can be assigned a conceptual label. We test our method on two scenes, using data acquired from in-the-wild web cams. Web camera data is challenging because it is often low resolution (spatially and temporally) with compression artifacts and noise.

Figure 1 contains an illustration of our learning and classification process. Our approach adapts the bag-of-words concept to trajectory-level behavioral analysis. First, on each video scene, tracks are computed for a period of time that is sufficient to capture the range of activity in the scene (typically a few hours or a day). We assume that the cameras are roughly calibrated to the ground plane, so that ground-plane tracking and normalization may be performed. Each video scene is partitioned into a set of regions, such as a regular spatial grid in the ground plane.

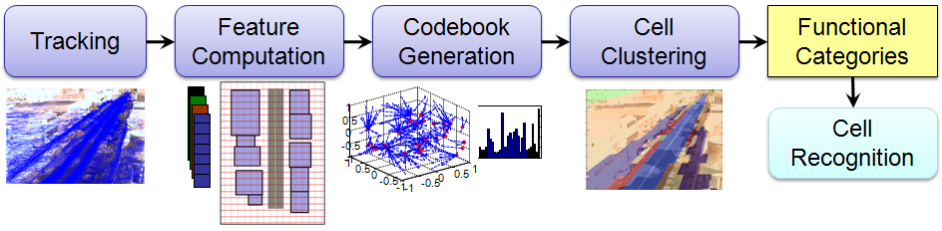


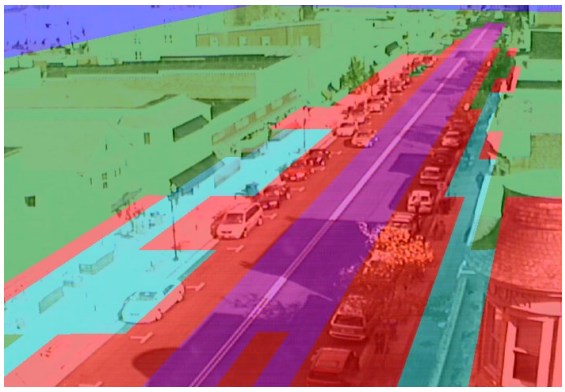
Figure 1. *Unsupervised modeling architecture.*

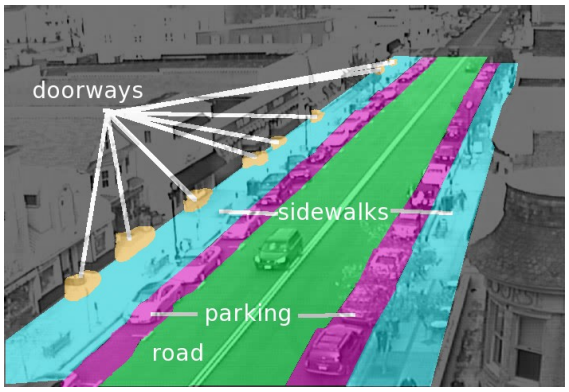
Next, a descriptive set of behavioral features is computed. For each track, and for each grid cell that the track intersects, low-level features capture single-object, local, behavioral and object characteristics such as velocity, heading change, speed change and size within the cell and in the nearby area. Within each cell, the low-level features are accumulated to form feature distributions in each cell. More significantly, mid-level features capture the relationship between cells traversed by the same track, and localized relationships between tracks over time. In total, the feature set characterizes local behaviors in the same way that patch descriptors characterize local appearance for object recognition. Our features are not tied to specific scene locations, allowing us to build models that generalize across scenes.

The feature vectors are clustered using mean-shift (or K-means) to form a codebook of a size that is comparable to the number of cells. For each cell on the ground plane, a codebook histogram is formed by finding the closest centroid for each feature vector in the cell. Cells are then clustered using mean-shift with the cell codebook histogram as the feature vector.

Figure 2 and Figure 3 contain results on two web-cam scenes including ground truth and PCC tables. In the first scene, Ocean City, we trained the functional scene element models using data from the Ocean City scene. We then classified the scene elements in that scene using the learned models. The second scene, Ware, was classified using the scene element models learned from the Ocean City scene. The models capture background, parking, and road regions particularly well. The PCC for road on Ware (0.75) is nearly as good as Ocean City (0.78). The sidewalk results on Ware are actually better (0.475) than Ocean City (0.3571). Parking results are good on Ocean City, but did not transfer well to Ware. Doorways were poor on both scenes. We plan to add additional features which capture acceleration and

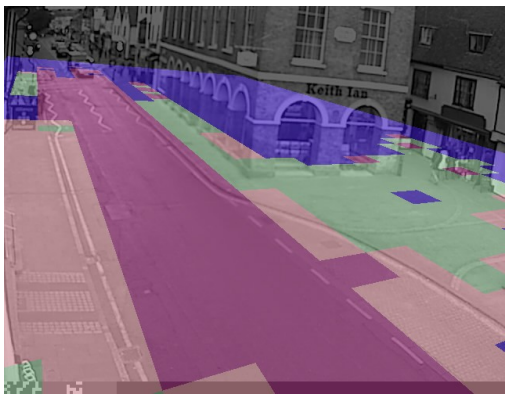
deceleration at the ends of tracks to help improve the classification of both doorways and parking areas. We are encouraged by these initial results, especially since several categories were learned on one scene and generalized well to another scene.

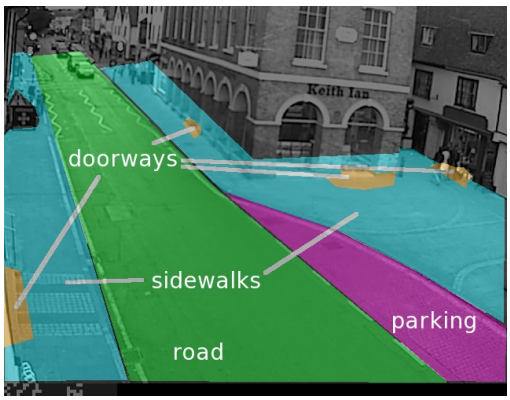




Class	Background	Doorway	Parking	Road	Sidewalk	PCC
Background	621	0	20	19	7	0.931
Doorway	2	0	0	0	2	0
Parking	1	0	31	0	1	0.9394
Road	0	0	11	39	0	0.78
Sidewalk	7	0	20	0	15	0.3571

Figure 2. *Unsupervised functional recognition results on Ocean City.*





Class	Background	Doorway	Parking	Road	Sidewalk	PCC
Background	391	0	0	4	14	0.956
Doorway	0	0	0	1	7	0
Parking	1	0	0	1	7	0
Road	37	0	0	126	5	0.75
Sidewalk	119	0	0	7	114	0.475

Figure 3. *Unsupervised results on the Ware web camera using functional element models learned on Ocean City.*

1. W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(9):1450–1464, 2006.
2. D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, pages 397–408, 2005.
3. C. Stauffer. Estimating tracking sources and sinks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2003.
4. X.Wang, K. T. Ma, G.-W. Ng, andW. E. L. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.