Pyramid Coding for Functional Scene Element Recognition in Video Scenes

Eran Swears[1], Anthony Hoogs[1], and Kim Boyer[2]

[1]Kitware Inc., {eran.swears|anthony.hoogs}@kitware.com
[2]ECSE Department, Rensselaer Polytechnic Institute, kim@ecse.rpi.edu

1. Introduction

We present a new approach to video scene modeling and scene element recognition that makes several improvements over existing state-of-the-art approaches. More specifically, we recognize stationary scene elements in video using descriptors derived from the moving objects (people/vehicles) that interact with them. When these scene elements have a specific purpose or function they are referred to as functional scene elements [2,3,4]. Some examples include: parking-spot, sidewalk, building, doorway, and cross-walk. Relying on descriptors derived from automatically computed tracks, as opposed to pixel features, enables the detection of scene elements that cannot be discriminated based on appearance alone. For example, cross-walks may or may not have the black and white zebra patterns and doorways can be completely occluded in high altitude aerial video as is the case with one of the video datasets analyzed here (Figure 1). Fortunately, the moving objects that interact with them are easier to detect and track which enables the detection of these visually ambiguous or poorly seen elements.

Existing approaches have a limited ability to characterize elements such as cross-walks, intersections, and buildings that are multi-modal, have low activity, or have indirect evidence. Multi-modal scene elements have multiple modes of behavior characteristics; for example, roadways have vehicles driving on them but they can also have vehicles stopping and turning to enter a parking-spot. Low activity scene elements have very few moving objects associated with them, while elements with indirect activity (building) have moving objects that are nearby, but not within their bounds. For example, a person can enter the building, but from an aerial perspective the building does not have movers on its roof.

Our solution for recognizing scene elements with multi-model activity is to introduce a pyramid coding approach that creates a hierarchy of descriptive clusters to form a pyramid that is sparse in the number of clusters and dense in content. Our first contribution is the characterization of scene elements using this sparse-dense pyramid of codebooks, which implicitly captures all behavior granularities, up to a maximum number, to enable the characterization of multi-modal behaviors.

Our second contribution is the incorporation of local behavioral context (person-enter-building, vehicle-parking-nearby) to compensate for both the low and indirect activity. Local behavioral context is captured by aggregating (pooling) behaviors that surround the scene element of interest, not just a single grid cell [2,3]. This increases the observed amount of activity and couples the scene element with nearby activity.

These two contributions significantly improve scene element recognition when compared against state-of-the-art approaches [1,2,3]. Further details on the approach and more results are discussed in our conference paper [4].

2. Approach

Our overall approach recognizes spatial regions that have similar functional behaviors as the presented training examples. Our framework for this involves a coding step and a pooling step. The coding step starts with a set of descriptors derived from moving objects, which are then fed into the pyramid coding algorithms that use hierarchical divisive clustering based on Gaussian Mixture Models (GMMs) to form the pyramid of codebooks.

After pyramid coding, a 2D spatial grid is applied to the scene's ground plane and encoded once for each codebook in the pyramid of codebooks. The scene element models are formed during the pooling step, where one model is created for each training example. Pooling involves accumulating the unique clusters for the Regions of

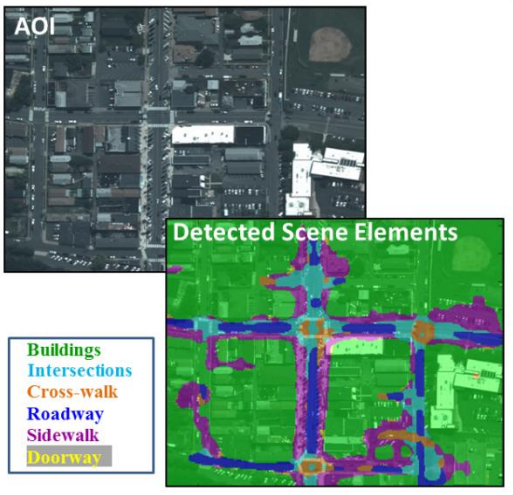

Figure 1, (Top) AOI from ARGUS-IS collected data around main street, (Bottom) detected functional scene elements.

Interest (ROIs) from each layer's encoded scene into a histogram model.

The testing process is a recognition framework that identifies both the location and label of scene elements. During the testing process an "unknown" histogram model from a test ROI is compared to each learned model which returns the likelihood of fitting to each. The scene is raster scanned with the test ROI to produce a 2D likelihood map that is later smoothed with a Markov Random Field (MRF) to enforce spatial class adjacency constraints.

3. Pyramid Coding

The pyramid coding process first forms the sparse-dense pyramid of codebooks and then encodes the scene into a pyramid of functional region maps [5]. The pyramid of codebooks is formed by clustering the set of descriptors through a hierarchical divisive clustering algorithm, which bifurcates the largest variance cluster in each layer to produce two unique clusters per layer. A variety of behavior characteristics are captured through the various descriptors types used during clustering such as events, classification results, and normalcy results.

Another benefit of this approach is that the clusters are dense in content, since they are split based on variance, and as such very few clusters need to be created to characterize complex scene elements. In fact, our pyramid only contains $2(L-1)$ unique clusters for L layers, while a comparable method such as the Hierarchical K-Means [1] that splits every cluster into K more clusters over L layers, contains $\sum_{l=1}^{L} K^l$ clusters that are sparse in content.

The codebooks at each layer are used to encode the scene, once per layer. This is accomplished by overlaying a $I \times J$ grid onto the ground plane and assigning each grid cell the most frequently occurring codeword for the layer of interest, which captures the normal behavior associated with a grid cell. The final result of the encoding process is a pyramid of functional region maps.

4. Modeling Local Behavioral Context

Local behavioral context is modeled by performing the pooling step on the training example's bounding region for each encoded layer of the pyramid. Average-pooling accumulates the number of times that the two unique clusters (mixture of behaviors) occur in each layer into a two bin histogram, which is then normalized by the size of the region. The histograms from each layer are concatenated to form the model, which is compared against the histogram extracted from a test window using the Laplace kernel as part of the testing process.

5. Experiments

Experiments are performed on the ARGUS-IS collected WAMI data [4], which is captured from a moving aerial platform and includes: Building (25), Intersection (17), Cross-Walk (18), Roadway (23), Sidewalk (50), and Doorway (11). Our approach is compared against the "Functional-Category" [2], and "Supervised-MRF" [3] functional recognition techniques using the mean Average Precision (mAP) and Probability of Correct Classification (PCC), see Table 1. Figure 1, shows the qualitative results of our approach, where the decision labels of MRF inference are colored coded for each scene element type.

Table 1, mAPs and PCCs on the WAMI data

Legend	Functional Category [2]		Supervised MRF[3]		Pyramid Coding [4]	
FE Names	mAP / PCC%		mAP / PCC%		mAP / PCC%	
Building	NA	0.6	0.06	0.02	0.92	79.6
Intersection	NA	18.3	0.50	0	0.58	52.2
Cross-walk	NA	4.5	0.15	18.7	0.21	32.4
Roadway	NA	50.3	0.42	66.1	0.63	44.3
Sidewalk	NA	34.9	0.46	66.0	0.53	65.0
Doorway	NA	5.9	0.01	0	0.03	0
Overall	NA	9.0	0.19	16.8	0.72	68.0

The Functional-Category method is an unsupervised mean-shift clustering algorithm while the "Supervised MRF" method uses manually defined functional scene element models based on binary descriptors combined with an MRF for smoothing.

The lower performance of the Functional-Category approach is due to the algorithm's sensitivity to the grid cell size, mean shift parameter, and because it does not model local context or multi-modal behavior. The Supervised MRF based method detected the Roadways and Sidewalks well, as seen by the PCCs in Table 1, while the others appear to have very low performance. We believe the lower overall performance here is also due to the fact that it does not capture the local behavioral context or multi-modal behavior.

6. Acknowledgement/Disclaimer

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract no. W91CRB-10-C-0098. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of DARPA or the U.S. Government. Approved for public release; distribution unlimited.

7. References

[1] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," CVPR, 2006

[2] M. Turek, A. Hoogs, and R. Collins, "Unsupervised Learning of Functional Categories in Video Scenes", ECCV, 2010

[3] C. Fernandez, J. Gonzalez, and X. Roca, "Automatic Learning of Background Semantics in Generic Surveilled Scenes," ECCV, 2010

[4] E. Swears, A. Hoogs, and K. Boyer, "Pyramid Coding for Functional Scene Element Recognition in Video Scenes," ICCV, 2013