Segmentation and Scene Modeling for MIL-based Target Localization*

Karthik Sankaranarayanan
IBM Research, India
kartsank@in.ibm.com

James W. Davis
Ohio State University, USA
jwdavis@cse.ohio-state.edu

Abstract

Existing techniques for object tracking with Multiple Instance Learning take the approach of extracting low-level patches of fixed size and aspect ratios within each image, and employ many simplistic assumptions. In this work, we propose an approach that automatically utilizes image segments as input primitives to develop a multi-level segmentation-based system, and build a target model refinement procedure that learns the optimal model corresponding to the target object. To go beyond existing restrictive assumptions, we further develop automatic scene environmental models to assign prior probabilities to segment instances of belonging to the target vs scene. We demonstrate impressive qualitative and quantitative results with tracking sequences in typical outdoor surveillance settings.

1 Introduction

Object tracking using Multiple Instance Learning (MIL) [1, 4] has become a popular approach primarily since MIL allows the flexibility of ambiguous labeling of target objects at the level of images in a video sequence instead of the actual individual locations within the image. But these existing techniques which take the approach of extracting low-level patches of fixed size and aspect ratio are approximate, need manual tuning, and implicitly assume a particular camera viewpoint and target pose. To move beyond these simplistic assumptions, especially with PTZ cameras having a wide-range of possible viewpoints and poses, we propose to utilize image segments as input primitives thereby providing a semantically more meaningful set of descriptors as instances. Further, to go beyond the run-time assumptions of such techniques such as motion models or manual initialization, we exploit the scene information by building one-time environment models which we then employ to automatically assign prior probabilities to instances as belonging to the target/foreground or scene/background.

*Appears in *21st International Conference on Pattern Recognition (ICPR)*, Tsukuba, Japan, Nov 2012.

The work presented here seeks to build on and improve the approach of [4] to detect a persistent target in a video sequence, which proposed a MIL optimization algorithm using softmax where the logistic function weight vector $\mathbf{w}$ models the target of interest (one for each target in case of multiple targets) and each instance x_i in a bag is represented using a vector of log-covariance features built from that instance patch.

2 Multi-level Segmentation Approach

We develop a multi-level segmentation approach based on the work of [2] by modeling the probability of a pixel being on a boundary conditioned on a set of locally measured image features such as Oriented Energy (OE) and Texture Gradient (TG).

$$OE_{\theta,s} = (I * f^e_{\theta,s})^2 + (I * f^o_{\theta,s})^2 \tag{1}$$

where $f^e_{\theta,s}$ and $f^o_{\theta,s}$ represent the quadrature pair of even and odd symmetric filters at the specified orientation θ and scale s, with f^e as Gaussian second derivative and f^o being its Hilbert transform. Using textons of [3], TG can be formally defined based on the χ^2 distance between the two histograms m and n of textons in the two disc halves around a pixel as

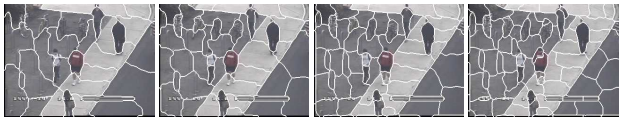
$$\chi^2(m,n) = \frac{1}{2}\sum \frac{(m_i - n_i)^2}{m_i + n_i} \tag{2}$$

The raw OE and TG signals are transformed to emphasize local maxima. Given a feature $f(x)$ defined over spatial coordinate orthogonal to the edge orientation, the derived feature is calculated as

$$\hat{f}(x) = f(x) \cdot \left(\frac{-f''(x)}{|f'(x)| + \epsilon} \right) \tag{3}$$

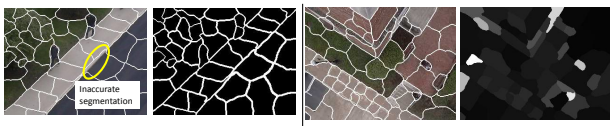
To robustly estimate the directional derivatives and localize the peaks, a cylindrical parabola is fit over a circular window of radius r centered at each pixel. The coefficients of the quadratic fit $ax^2 + bx + c$ directly provide the signal derivatives, so the transform becomes

$$\hat{f} = -\frac{2 \cdot c \cdot a}{|b| + \epsilon} \tag{4}$$



$n = 30$ $n = 50$ $n = 70$ $n = 90$

Figure 1. Hierarchy of segmentations (# of segments n from 30 to 100 in steps of 20).



(a) (b)

Figure 2. (a) Size-adaptive segment erosion. (b) Noisy segments have high intensity variance

and this transformation is applied to the OE and TG values at each θ and s separately. Such a combination of OE and TG allows us to more robustly detect natural boundaries.

An artifact of most segmentation algorithms is that the segmentation is not always accurate along the segment borders, resulting in the some of the pixels from the surrounding segments creeping into a segment (see Fig. 2(a)). To reduce this "structural noise" along segmentation borders, we perform an erosion operation on each segment adaptively based on the actual size of the segment (as for smaller segments we do not want to discard too many pixels, whereas for larger segments we can afford to remove more pixels). We then build upon the MIL framework of [4], where for each image in the training sequence of tracking, we run this segmentation over multiple levels, and then employ these segments as instances to form the input bags for the optimization algorithm to learn persistent targets across the bags.

3 Scene Modeling Approach

The next step in our approach is to exploit the scene structure to assign probability values to individual segments as belonging to scene/background or target/foreground. We develop a 3-step algorithm for this.

3.1 Scene Entity Probability: Barring actual changes in the structure of the scene (buildings, street, etc.) or environmental changes (e.g. seasonal change), the nature of the scene remains fairly similar. To exploit this property, we seek to model the environment-level structural information obtained from multiple views of PTZ cameras overlooking the entire scene (this is different from a single view-specific image-level background model, typically obtained in pixels).

For building such a scene model, we employ segments from across the entire scene, by collecting image data from the PTZ cameras from random pan-tilt locations and then run the segmentation algorithm on each of the images. We then build a Gaussian Mixture Model of segments with YCbCr features using the Expectation-Maximization (EM) algorithm and automatically select from different models, the model that maximizes the Bayesian Information Criterion (BIC) [5]. As seen in Fig. 3, the main clusters emerging from the scene are those corresponding to environmental features such as streets, walkways, buildings, and vegetation/grass.

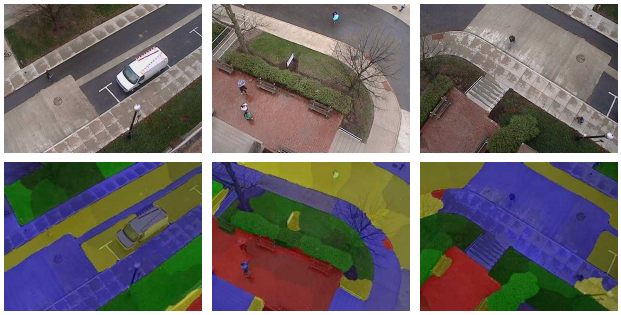


Figure 3. Nearest neighbor classification of segments. Color denotes the cluster assigned and the intensity of color corresponds to the probability value of belonging to that cluster.

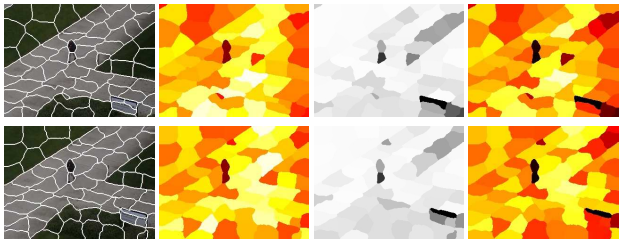
Next, using the GMM, each segment in a new image is assigned a background cluster using a nearest neighbor classification scheme in the YCbCr space with the Mahalanobis distance, and the Gaussian pdf (μ,σ) corresponding to that cluster is used to obtain the probability that the segment belongs to its assigned background cluster, as given by Eqn. 5

$$p_i^{gaussian} = \exp\left(\frac{-(x_{seg_i} - \mu_k)^2}{2\sigma_k^2}\right) \quad (5)$$

where x_{seg_i} is the feature vector corresponding to the ith segment seg_i, and the mean and standard deviation of its closest cluster k are μ_k and σ_k respectively.

3.2 Variance-based Weighting: The segmentation algorithm generally results in segments that are mostly uniform in color and hence any segment that contains a large variance in color is most likely a result of incorrect segmentation (mix of foreground-background, or segment contains more than one person, see Fig. 2(b). Therefore, we look at the intensity variance within each segment and employ a Laplacian-based weighting factor on the scene priors, as given in Eqn. 6

$$\beta_i = \exp\left(\frac{-\sigma_{seg_i}^2}{\sigma_n}\right) \quad (6)$$



(a) Image (b) Gaussian probability (c) Variance weighting (d) Final probability

Figure 4. Two step weighted-probability assignment.

where β_i is the variance-based weight assigned to segment i whose intensity variance is $\sigma^2_{seg_i}$, and σ_n is a standard scaling parameter for the Laplacian-based weighting scheme that is manually specified.

3.3 Combined Scene Prior Probability: These variance-based weights β_i are used along with Gaussian scene probability values $p_i^{gaussian}$ to obtain a weighted probability for each segment as

$$p_i^{bg} = \beta_i \cdot p_i^{gaussian} \tag{7}$$

Therefore, the final probability of a segment belonging to the scene (desired high) is composed of two parts, (i) Gaussian probability of belonging to a scene entity (desired high), and (ii) variance-weighting indicating uniform segment and hence likely to emerge from a scene entity (desired high) - see Fig. 4.

3.4 MIL Initialization: These combined scene probability values p_i^{bg} are used as labeling information on the segments to obtain a robust prior w_0 for the MIL optimization of [4]. We do this using a Maximum Likelihood Estimation approach, so that given the segmentation data (using its covariance features) with the background prior information $(\mathbf{X^{seg}}, \mathbf{p^{bg}}) = (x_1, p_1^{bg}), ..., (x_N, p_N^{bg})$, we find the weight vector w which maximizes Eqn. 8

$$w_0 = \arg\max_w - \sum_{i=1}^{N} \log\left(1 + \exp(-p_i^{bg} w^T x_i)\right) \tag{8}$$

which is optimized using gradient descent. The next part describes how we handle multiple disjoint, adjacent, and overlapping models learned from the sequence in our approach.

3.5 Refinement Procedure: In order to handle the possibility that multiple concepts could be learned from just one semantically persistent target (see Fig. 5(b) and (c)), we develop a refinement procedure based on the sizes, adjacency and overlap of the learned target concepts so as to retain only the concept that is the largest and a superset of the other learned concepts - see Alg. 1. The worst-case time complexity of this algorithm is $O(n^2)$ where n is the number of segments in each bag. Following refinement, the target model corresponding to the largest superset segment is retained as the most persistent concept (see Fig. 5(a)) and employed for performing tracking-by-detection in every successive frame in the video sequence.

Input: Representative set S_R of segments s with size (in number of pixels) n and centroid c of each.

Sort the segments of S_R based on size n
`/*Initialize each segment as not a sub-segment*/`
For each segment $s \in S_R$ set $s^{sub} = 0$;
Initialize $i = |S_R|$

repeat iterate i

  repeat iterate j
    Check if centroid $c_j \in s_i$
    If yes, $s_j^{sub} = 1$; `/*Mark segment` s_j `as sub-segment*/`
  until $j \leq i$;

  `/*Check if all segments in` S_R `are marked as sub-segments*/`
  For each $s \in S_R, s \neq s_i$, if $s^{sub} = 1$, break;
  Else $i = i - 1$;

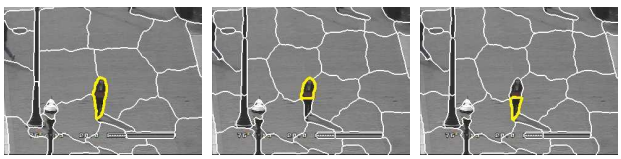
until $i \geq 1$;

$S_{fin} = \{s \in S_R, s^{sub} = 0\}$
Output: S_{fin} set of segments that are the largest superset of the segments of S_R.

Algorithm 1: Target models refinement procedure using representative segments.

4 Experiments

5.1 Target Localization with Segmentation: For each image in the training sequence, we performed segmen-



(a) (b) (c)

Figure 5. Representative segments for refinement, (a) is learned as the largest superset segment.

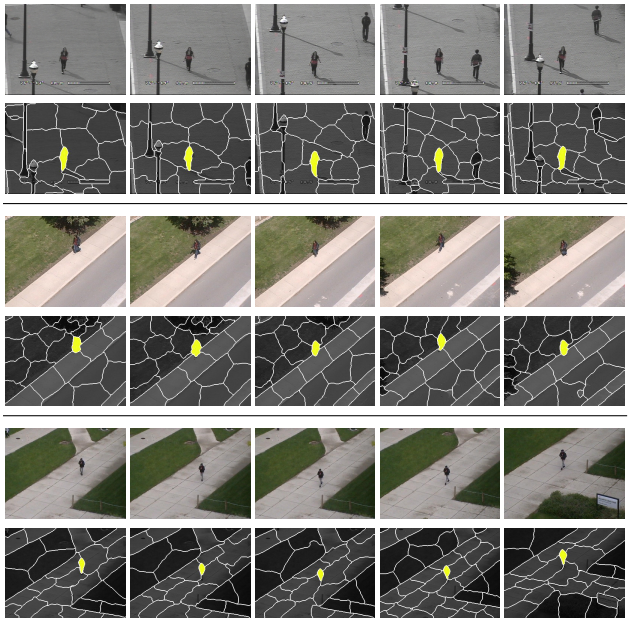
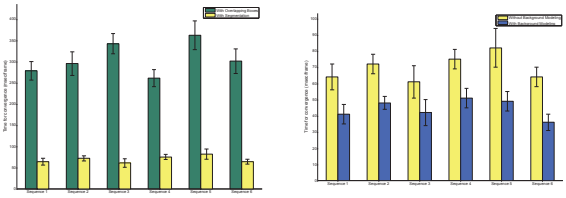


Figure 6. Target localization results on different video sequences using segments as input instances.



(a) (b)

Figure 7. Convergence speed with/without (a) using segments as instances, (b) employing scene priors.

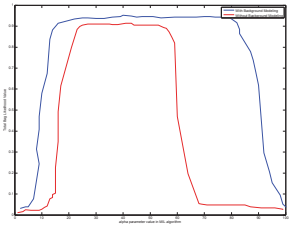


Figure 8. Stability: Overall likelihood of the learned model wrt parameter α with/without employing scene priors.

tation at 8 different levels, created positive and negative bags, using the priors to initialize the MIL algorithm as explained in Sect. 3, and performed the refinement procedure explained in Alg. 1 to obtain the largest superset segment and the corresponding model. In the testing phase, we tested this target model against each of these segments, with each such test gives us the probability of that segment representing the target of interest. Figure 6 shows the results for different video sequences where a probability heatmap is overlaid on the segmented testing images. From each of the above sequences, we can see that the algorithm is able to localize on the target well, thus validating the applicability of employing the segmentation enhancement for input instances to enhance the MIL approach.

5.2 Computational Efficiency: The plots in Fig. 7(a) demonstrate the speed-up achieved using segments as instances in the framework instead of technique of [4] by comparing the time for convergence (in msec/frame) on each of the 6 standard sequences from [4] with and without segmentation (see Fig. 7). We observed an overall speedup of 4-5 times in the running time of the MIL algorithm. Similarly, the plots in Fig. 7(b) demonstrates the speedup achieved by learning the scene model and using it to assign prior probabilities to image segments. From this, we conclude that a 35-40% improvement in the running time of the optimization procedure is achieved by employing scene priors within the MIL framework.

5.3 Parameter Stability with Scene Priors: The MIL algorithm [4] converges to the correct model only for a particular range of values for α parameter, outside of which a degenerate model is learned. This validity of convergence is checked by verifying the overall bag likelihood of the training data using the learned model. We observed that with scene prior probabilities, the range of values α that the algorithm permits for use was vastly increased (as seen in Fig. 8).

These experiments together validate our approach with PTZ camera tracking sequences in typical outdoor surveillance settings and demonstrate the speedup and stability achieved in the performance of the MIL algorithms by employing our work.

We gratefully acknowledge the support of the U.S. Department of Energy through the LANL/LDRD Program under LDRD-DR project RADIUS for this work.

References

[1] B. Babenko, M. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, 2009.

[2] M. D., F. C., and J. Malik. Learning to find brightness and texture boundaries in natural images. In *NIPS*, 2002.

[3] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *IJCV*, 2001.

[4] K. Sankaranarayanan and J. Davis. Object association across ptz cameras using logistic mil. In *CVPR*, 2011.

[5] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, pages 461–464, 1978.