Online Figure-ground Segmentation with Edge Pixel Classification

Zhaozheng Yin Robert T. Collins
Department of Computer Science and Engineering
The Pennsylvania State University, USA
{zyin,rcollins}@cse.psu.edu, http://vision.cse.psu.edu/

Abstract

The need for figure-ground segmentation in video arises in many vision problems like tracker initialization, accurate object shape representation and drift-free appearance model adaptation. This paper uses a 3D spatio-temporal Conditional Random Field (CRF) to combine different segmentation cues while enforcing temporal coherence. Without supervised parameter training, the weighting factors for different data potential functions in the CRF model are adapted online to reflect changes in object appearance and environment. To get an accurate boundary based on the 3D CRF segmentation result, edge pixels are classified into three classes: foreground, background and boundary. The final foreground region bitmask is constructed from the foreground and boundary edge pixels. The effectiveness of our approach is demonstrated on several airborne videos with large appearance change and heavy occlusion.

1 Introduction

Figure-ground segmentation is crucial for automated object detection and long-term object tracking under complex scenarios. For example, instead of being manually selected as a rectangular box in the first frame, an object should be segmented automatically from the background to initialize a tracker. Another example is long-term tracking, which requires adaptation to changes in object and background appearance while avoiding drift. The problem of drift arises because most trackers do not have a clear concept of object - their representation is a patch of pixels or a color histogram, with no explicit representation of figure or ground in the image. However, if we can explicitly segment the foreground from background, it would be possible to keep the adaptive model anchored on just the foreground pixels. In addition to solving the drift problem, an expected byproduct of segmentation for tracking is the extraction of more accurate and complete foreground bitmasks, leading to improvements in the analysis of object shape and detection of partial occlusion. A precise object representation is also helpful to search for and recognize the object again when a tracker loses that object.

In this paper, we solve the above problems by performing a precise figure-ground segmentation of moving objects using a Conditional Random Field model that combines multiple segmentation cues. Edges classified to be on the foreground object and its boundary further refine the accuracy of the foreground object bitmask.

1.1 Related Work

Image segmentation has been treated as a graph partition problem that simultaneously minimizes the cross segment connectivity and maximizes the within segment similarity of pixels [4, 16]. For video sequences, layered motion models have been one of the key paradigms for segmenting objects, assuming restricted parametric motion models or an elliptical shape model [17, 18]. Probabilistic fusion of multiple features has also been used successfully to segment objects [3, 6, 14]. Specifically, the Conditional Random Field (CRF [10, 9]) approach has been shown to be effective for combining different segmentation cues [3, 14].

When building a 2D CRF model for video segmentation, two unavoidable issues arise: how to maintain temporal coherence and how to estimate weighting parameters for different data potential functions. In [3], a second-order Markov chain is used as a temporal prior energy term to impose temporal continuity of graph nodes. In [14], the foreground and background are divided into triangular regions in each frame. After the correspondences of triangles between two frames are solved by linear programming, the temporal prior of any current triangle is computed as a weighted average of the previous triangles. Rather than computing temporal coherence as an extra prior energy term in a 2D spatial CRF model, we naturally extend the 2D CRF into a 3D CRF model to enforce spatio-temporal coherence.

The weighting parameters for different energy terms are usually estimated by maximizing the log-likelihood of the training data [10, 11]. This supervised parameter learning is good for static scenes having a comprehensive training data set. However, for a dynamic video sequence, it is more suitable to update the weights adaptively to reflect changes in appearance and environment. In [3, 14], the weighting parameters for different segmentation features are set by hand and remain the same during the whole video. In this paper, we propose an online parameter updating scheme for the 3D CRF model.

Since figure-ground segmentation aims to find an accurate boundary between foreground and background, edge pixels deserve more attention. For example, active contours (snakes [8]) and intelligent scissors [12] are two successful interactive tools to trace an object boundary - when seed points are chosen in proximity to the object edge, a minimum cost contour will snap to the object of interest. In this paper, we classify edge pixels into three categories based on the 3D CRF region segmentation. The edges belonging to the foreground and on the boundaries are used to generate the foreground bitmask, while the edges belonging to the background are ignored.

2 Figure-ground Segmentation

Originally introduced as a 1D sequential model in [10] and further extended to a 2D image lattice in [9], Conditional Random Fields (CRF) have been widely used to solve segmentation and discriminative labeling problems. Kumar and Hebert [9] also show that CRF offers several advantages over Markov Random Fields (MRF [5]): the MRF assumption of conditional independence of the data is removed in CRF, and the unary data association potential is defined over all the observations in CRF, rather than as a function of only local nodes as in MRF. Furthermore, in a MRF, the pairwise interaction potential between neighboring nodes only depends on node labels (i.e. this potential favors similar labels but penalizes dissimilar labels), while a CRF defines the interaction potential over the nodes AND all the observation data (e.g. allowing data-dependent interaction that

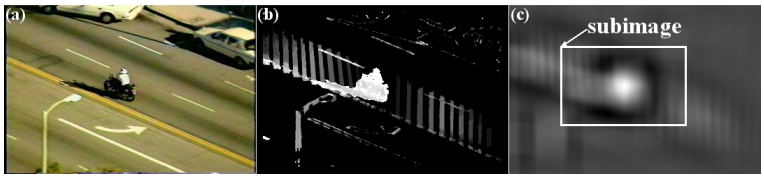

Figure 1: (a) Input image; (b) Motion detection by motion history images; (c) Mode and scale detection of the motion blob.

considers intensity/color discontinuity).

Let $\{I_i\}$ and $\{s_i\}$ denote the sets of image pixels and corresponding labels respectively. Label $s_i = 1$ if I_i belongs to the foreground, and $s_i = -1$ otherwise. Let $\{V,E\}$ be a graph such that s is indexed by the vertices V and E contains all the pairwise links between two neighboring nodes. Globally conditioned on the observation I, the joint distribution over labels s is

$$P(s|I) = \frac{1}{Z}\exp\left\{\sum_{i\in V}\sum_{k}\lambda_k\Phi_k(s_i,I) + \sum_{<i,j>\in E}\Psi(s_i,s_j,I)\right\} \tag{1}$$

where Z is a normalizing factor. The Φ_k's are data association potentials generated by different segmentation cues, and they are linearly combined with weights λ_k. Ψ represents pairwise interaction potentials between spatio-temporal neighboring nodes.

Data association potentials, Φ_k, measure how likely it is that node i has label s_i given image I without considering other nodes in the graph. Suppose $f_k(\cdot)$ is a function to map the input image to segmentation feature space such that $f_k : \Re^2 \rightarrow \Re$. The kth association potential at node i with label s_i is defined as

$$\Phi_k(s_i,I) = \log p(s_i|f_k(I)) \tag{2}$$

Similar to [3, 11], the interaction potential between nodes s_i and s_j given image I is defined as

$$\Psi(s_i,s_j,I) = s_i s_j \cdot e^{\frac{-\|I_i-I_j\|^2}{2\beta^2}} \tag{3}$$

where $\beta = <\| I_i - I_j \|^2>$ and $< \cdot >$ is the expectation operator. This potential describes how neighboring nodes interact. For example, if the pixel color difference is small ($|I_i - I_j| < \beta$), this potential encourages nodes s_i and s_j to have the same label.

2.1 Features for Segmentation

Segmenting figure from ground using a single feature alone can be expected to be error-prone, especially when nearby confusers are present. However, the CRF model is suitable for probabilistically fusing arbitrary, overlapping and agglomerative observations from both the past and future [10]. In this section, we introduce a suite of useful features for figure-ground segmentation in video.

First of all, motion detection is important to trigger the automatic segmentation process, and motion is also a powerful feature for moving object segmentation. To detect object motion from a moving camera, we adopt the forward/backward motion history image (MHI) method, which combines motion information over a sliding temporal window [20]. Figure 1(b) shows a sample motion detection result denoted as $f_{MHI}(\cdot)$. Because the detection result is noisy due to errors in camera motion compensation, parallax, and fluctuations of background appearance, we detect the mode location (x,y) and scale (σ_x,σ_y)

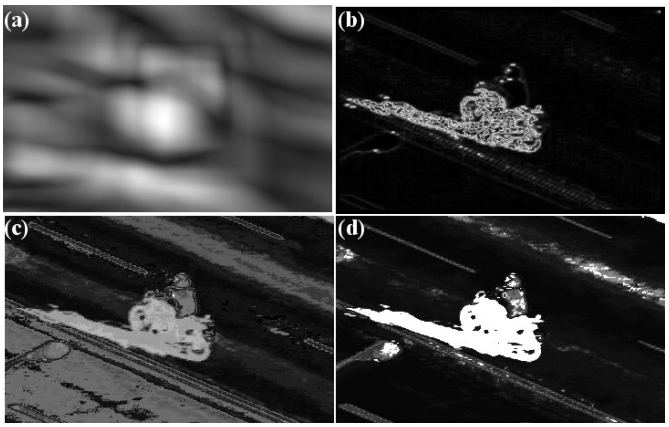

Figure 2: A suite of likelihood features for figure-ground segmentation: (a) center-surround saliency; (b) local color contrast; (c) spatial color variance; (d) figure-ground color likelihood ratio.

of the motion blob by using the mean-shift in scale space method [2]. Other segmentation features are then extracted from the subimage centered at the mode (Figure 1(c)) rather than the whole image to reduce computational cost.

In addition to $f_{MHI}(\cdot)$, we compute other segmentation features including center-surround saliency, local color contrast, spatial color variance and figure-ground color likelihood ratio (see Figure 2). Saliency-driven visual attention detects objects that "jump out" from their surroundings, i.e. an object distinct from its surrounding background deserves visual attention [7, 11]. We compute the center-surround color histogram distance as a measure of saliency. For each pixel location u in the image, we extract the color histogram, h_u^{IN}, within a rectangle centered at u with scale (σ_x, σ_y), and another histogram within the surrounding ring, h_u^{OUT}. The saliency value at u is computed by Earth Mover's Distance (EMD [15]), which is relatively insensitive to changes in intensity and color, as well as to the parameters of the histogram construction. We represent the color histogram by a set of 1D marginal distributions, and make use of the fact that the EMD distance can be computed in closed form as an L1 distance between 1D cdfs [1].

$$f_{CS}(I,u) = |\text{cdf}(h_u^{IN}) - \text{cdf}(h_u^{OUT})| \tag{4}$$

Finally, $f_{CS}(\cdot)$ is normalized to lie in the range $[0,1]$. We use the integral histogram method [13] to speed up the saliency feature computation.

Contrast has widely been used to capture local color difference [7, 11]. Here, we compute color contrast at pixel u as

$$f_{\text{CONTRAST}}(I,u) = \sum_{v \in N(u)} \frac{\| I_u - I_v \|^2}{\| I_u + I_v \|^2 + \varepsilon} \tag{5}$$

where $N(u)$ is a 5×5 local window around pixel u. The small ε is added to avoid dividing by zero. $f_{\text{CONTRAST}}(\cdot)$ is also normalized to the range $[0,1]$.

Color spatial distribution is a global feature related to object saliency. It is observed in [11] that colors distributed over a larger spatial area of the image are less likely to belong to a salient foreground object. Instead of modeling all colors in an image by Gaussian mixture models as done in [11], we directly compute a color's x-coordinate variance as

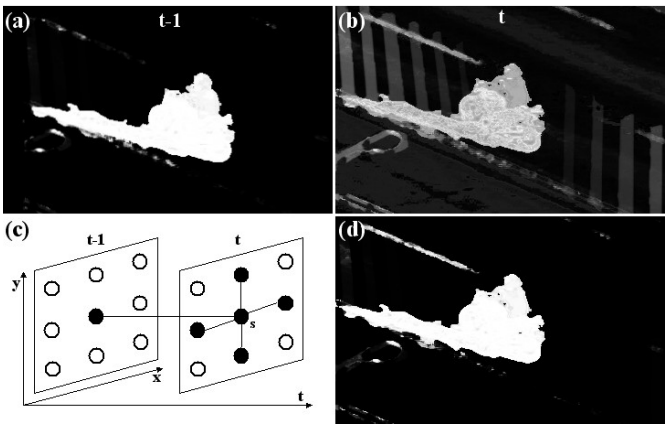

Figure 3: (a) Inference result in the previous frame (transformed into the current frame's coordinate system); (b) Probabilistic fusion of segmentation features in the current frame; (c) 3D CRF model; (d) Inference result in the current frame.

$$\text{var}_x(c) = \frac{\sum_v \delta(I_v = c)(v_x - m_x(c))^2)}{\sum_v \delta(I_v = c) + \varepsilon} \tag{6}$$

where $\delta(\cdot)$ is an indicator function, v_x represents the x-coordinate of pixel v, and $m_x(c)$ is the x-coordinate mean of those pixels with color c, computed as:

$$m_x(c) = \frac{\sum_v \delta(I_v = c) v_x}{\sum_v \delta(I_v = c) + \varepsilon} \tag{7}$$

Similarly, we compute a color's y-coordinate variance for $\text{var}_y(c)$. Although the traditional spatial variance of a color component is a 2-by-2 covariance matrix, here we approximate it by a scalar $\text{var}(c) = \text{var}_x(c) + \text{var}_y(c)$. After normalizing $\text{var}(\cdot)$ to the range $[0,1]$, the color spatial-distribution at pixel u is defined as

$$f_{VAR}(I,u) = 1 - var(I_u) \tag{8}$$

Finally, given a foreground bitmask from the previous frame, we can extract an accurate foreground color appearance model, e.g. color histogram H_F. The background appearance model, H_B, is computed from the complement of the foreground bitmask. The figure-ground color likelihood ratio at pixel u is then computed as

$$f_{CLR}(I,u) = \frac{H_F(I_u)}{H_B(I_u) + \varepsilon} \tag{9}$$

Note that $f_{CLR}(\cdot)$ is only available after the first frame, because it relies on a previous segmentation result.

2.2 Label Inference

We add a temporal coherence link into the popular 2D spatial CRF model to obtain the 3D CRF model shown in Figure 3(c). Each node s_i in the current frame thus has four 4-connected spatial neighbors and one temporal neighbor. The five links are weighted based on how similar the pixels are (Eq.3). Since the camera is moving, the previous segmentation result is first transformed into the current frame's coordinate system (Figure 3(a)) using a warping matrix estimated via image stabilization. Based on all the data

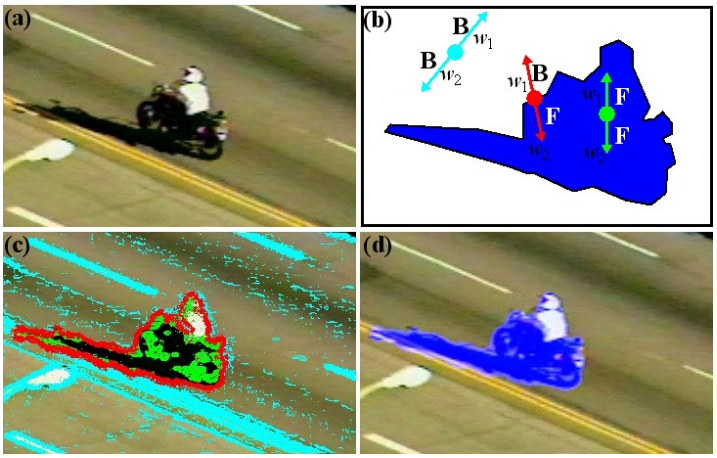

Figure 4: (a) Input image; (b) Feature extraction along the normal direction of edge pixels; (c) Edge pixel classification: figure-ground boundary (red), foreground edges (green), background edges (cyan); (d) Foreground mask (blue). Better seen in color.

association potentials (Figure 3(b)) and the spatial-temporal interaction potential, we run loopy belief propagation (BP [19]) to solve for s_i at each pixel. Since the graph $\{V,E\}$ has a regular grid structure, we perform asynchronous accelerated message updating. Four 1-dimensional BP sweeps in the spatial domain (left to right, up to down, right to left and down to up) are performed individually and in parallel. During each iteration, a node's belief is computed from five incoming messages (four from spatial neighbors and one from the temporal neighbor). The inference process converges very fast, typically in less than 10 iterations. Figure 3(d) shows an example of the final inference result.

2.3 Edge Pixel Classification

Observing that the inference result from the 3D CRF is not perfect (Figure 3(d)), to get a more accurate figure-ground segmentation boundary, we draw attention to the salient edge pixels. The spatial gradient at pixel u is computed by first order Gaussian derivative and denoted as $\vec{g}(u) = (g_x(u), g_y(u))$ with magnitude $||g(u)||$ and direction $\vec{n}(u) = (\cos(\theta(u)), \sin(\theta(u)))$ where $\theta(u) = \tan^{-1}(g_y(u)/g_x(u))$. After normalizing all the pixel gradient magnitudes to the range $[0,1]$, we consider pixel u as an edge pixel if $||g(u)|| > 0.1$, and extract feature vector $\vec{w}_1(u)$ along $\vec{n}(u)$ as

$$\vec{w}_1(u) = \{p(s_v = 1|I), \forall v,\ s.t.\ |((u_x,u_y)-(v_x,v_y)) \times \vec{n}(u)| = 0 \& ||(u_x,u_y)-(v_x,v_y)|| \leq L\} \tag{10}$$

where '$\times$' is the cross-product between two vectors. This essentially samples L pixels along a line in the direction of the gradient vector, i.e. to one side of the edge boundary. We choose the feature vector length $L = 10$. Similarly, we get $\vec{w}_2(u)$ along $-\vec{n}(u)$. Edge pixel u is classified as a boundary pixel between foreground and background if

$$\min(\bar{w}_1(u), \bar{w}_2(u)) < 0.5 \ \&\ \max(\bar{w}_1(u), \bar{w}_2(u)) > 0.5 \tag{11}$$

where $\bar{w}_1(u)$ and $\bar{w}_2(u)$ represent the mean of the feature vector. Although a more complicated feature distance measure can be applied to $\vec{w}_1(u)$ and $\vec{w}_2(u)$, this simple classifier works well on our datasets. The intuition is that one of the two feature vectors belongs

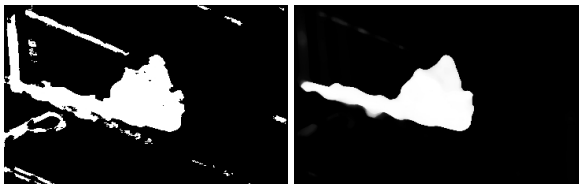

Figure 5: Two alternative approaches, (a) Thresholded result of the probabilistic fusion and (b) Median filter on the 3D CRF inference result, yield inferior foreground bitmask results as compared to our proposed method (Compare to Figure 4(d)).

primarily to foreground and the other one belongs primarily to background. Similarly, we classify other edge pixels as belonging completely to foreground or completely to background. Figure 4(c) shows an example of these three classes of edge pixels. Finally, all the foreground and boundary pixels are morphologically connected and the inside hole is filled to form a foreground bitmask (Figure 4(d)).

2.4 Online Parameter Tuning

The parameter estimation problem in the 3D CRF is to determine the weighting factors (λ_k) for different potential functions. For long video sequences, it is tedious to get a complete ground truth data set for training. Furthermore, it is desirable to dynamically update the parameters over time to adapt to scene and lighting changes. We apply discriminant analysis on each segmentation cue to measure its ability to discriminate between figure and ground. This is motivated by the fact that the segmentation features with high foreground/background discrimination deserve high weight so they can make a significant contribution to the segmentation in the next frame. For each feature map, two data clusters are extracted

$$Z_F = \{f_k(I,u) : u \in F\};\;\; Z_B = \{f_k(I,u) : u \in B\} \tag{12}$$

where F and B are bitmasks for foreground and background as shown in Figure 4(b). A straightforward measure of the separation between clusters Z_F and Z_B is the difference of sample means. However, if the sample mean difference is large but the data distributions of the two clusters are multi-modal or highly overlapped with each other, the two clusters are not well separated. Thus, we examine the sample mean difference relative to some measure of the standard deviations. Inspired by Fisher linear discriminant analysis, we define the figure-ground separability, i.e. weight λ_k, as

$$\lambda_k = \max(0, \frac{\bar{Z}_F - \bar{Z}_B}{\text{std}(Z_F) + \text{std}(Z_B)}) \tag{13}$$

where $\bar{Z}_F$ and $\text{std}(Z_F)$ represent the mean and standard deviation of cluster Z_F respectively, $\lambda_k > 0$ and we normalize such that $\sum_k \lambda_k = 1$. The computed weights in the current frame are applied to the next frame. In the first frame, the weight for f_{CLR} is zero since F and B are not available yet, while other segmentation features have equal initial weights.

3 Experimental Results

Figure 5(a) shows a simple thresholded mask of Figure 3(b) formed from the probabilistic fusion of different segmentation features. The hard-thresholding method causes many

mis-segmented pixels. Applying median filtering on the 3D CRF inference result (Figure 3(d)) removes some outliers but also loses fine detail on the foreground/background boundary. In comparison, our approach based on classifying edge pixels preserves fine detail to delineate an accurate boundary (Figure 4(d)).

We have tested our approach on several airborne videos with low color quality. Objects change their shape and appearance throughout the videos, and often undergo partial or full occlusion. All the segmentation processes are automatically started by motion detection. Every object maintains its own foreground mask and appearance representation. When there are no detected motion blobs in the predicted location of an object, it is considered occluded, and we predict its trajectory. When the occluded objects are detected again by motion detection, we use nearest-neighbor data association to decide which recovered object corresponds to the predicted object location.

Figures 6 and 7 show some example segmentation results with promising performance. Videos with thousands of result frames have been submitted as supplemental material. Aithough our method works well much of the time, we want to point out some of the misclassified edge pixels in our experiments. For example, when an object image is split into two halves by a small pole (Figure 6(c)), the edge pixels around the pole are incorrectly classified. In addition, when the color of pavement markings fluctuates in the video (Figure 6(e)), they are detected as a motion blob and misclassified as foreground. We believe further research on object shape representation and appearance modeling can help improve the current results. For example, we only use a bottom-up approach to classify edge pixels, while current figure-ground performance could be improved if a top-down object model is explored (e.g. shape constrained edge classification).

4 Conclusion

In this paper, we propose an approach for figure-ground segmentation of moving objects in video sequences. A 3D CRF model is applied to combine different features and maintain coherence between temporal neighboring nodes. The weighting factors for different data potential functions are updated online to adapt to complex scenarios. To obtain accurate boundary information between foreground and background, we classify salient edge pixels into three classes: (1) within FG, (2) within BG, and (3) between FG and BG. The foreground and boundary edge pixels are then used to form the foreground object bitmask. This automatically extracted segmentation mask has applications to tracker initialization, drift-free object model adaption, and shape analysis. The approach has been tested on several airborne video sequences with large appearance change and occlusion, and yields promising results.

References

[1] S. Cohen, "Finding color and shape patterns in images," Technical Report STAN-CS-TR-99-1620, Stanford University, May 1999.

[2] R. Collins, "Mean-shift Blob Tracking through Scale Space," In CVPR, p234-240, 2003.

[3] A. Criminisi, G. Cross, A. Blake and V. Kolmogorov, "Bilayer Segmentation of Live Video," In CVPR, p53-60, 2006.

[4] P. Felzenszwalb and D. Huttenlocher, "Efficient Graph-Based Image Segmentation," Int'l. J. Comp. Vision, 59(2):167-181, 2004.

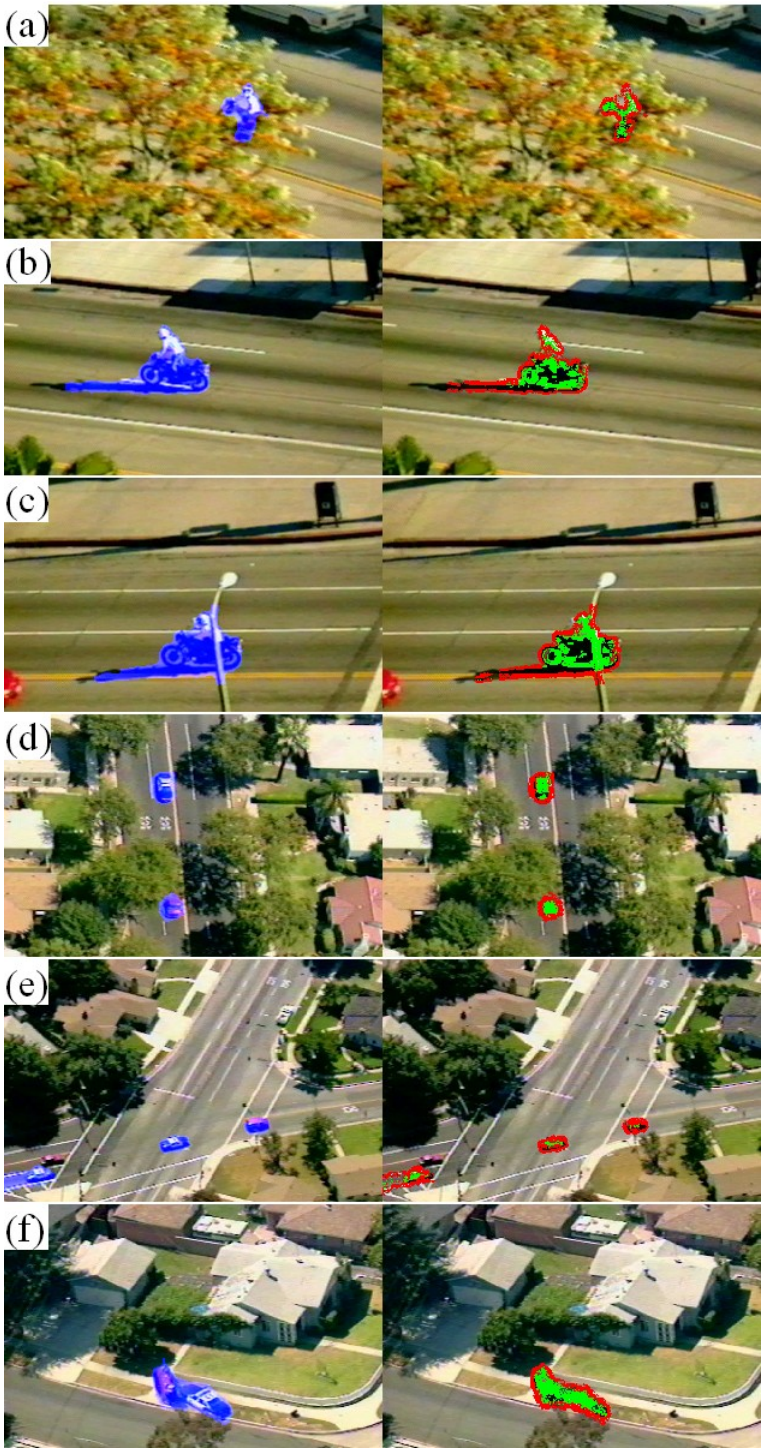

Figure 6: Left column: input video with foreground mask (blue); Right column: foreground edge pixels (green) and boundary edge pixels (red). Video demos are online.

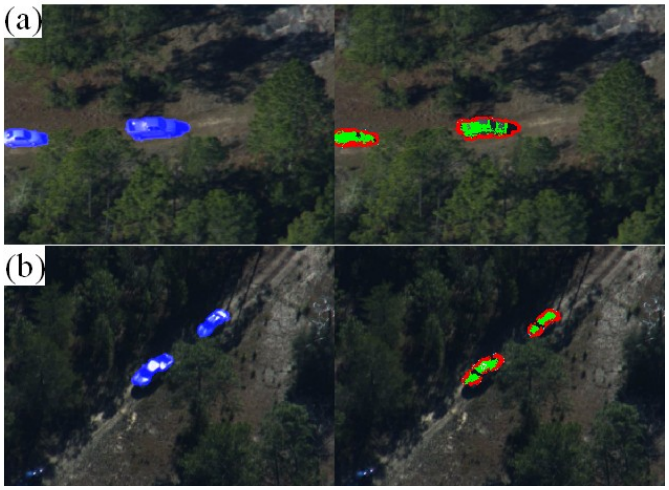

Figure 7: Left column: input video with foreground mask (blue); Right column: foreground edge pixels (green) and boundary edge pixels (red). Video demo is online.

[5] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images," IEEE Trans. PAMI, 6:721-741, 1984

[6] E. Hayman and J. Eklundh, "Probabilistic and Voting Approaches to Cue Integration for Figure-Ground Segmentation," In ECCV, 2002.

[7] L. Itti, C. Koch and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," IEEE Trans. PAMI, 20(11): 1254-1259, 1998.

[8] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active Contour Models," Int'l. J. Comp. Vision, 321-331, 1988.

[9] S. Kumar and M. Hebert, "Discriminative Random Fields," Int'l. J. Comp. Vision, 68(2):179-201, 2006.

[10] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," In ICML, 2001

[11] T. Liu, J. Sun, N. Zheng, X. Tang and H. Shum, "Learning to Detect A Salient Object," In CVPR, 2007.

[12] E. N. Mortensen and W. A. Barrett, "Intelligent Scissors for Image Composition," In SIGGRAPH, pp. 191-198, 1995.

[13] F. Porikli, "Integral Histogram: A Fast Way to Extract Histograms in Cartesian Spaces," In CVPR, p829-836, 2005.

[14] X. Ren and J. Malik, "Tracking as Repeated Figure/Ground Segmentation," In CVPR, 2007.

[15] Y. Rubner, C. Tomasi and L. Guibas, "The Earth Mover's distance as a Metric for Image Retrieval," Int. J. of Computer Vision, 40(2), 99-121, 2000.

[16] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," IEEE Trans. PAMI, 22(8), 888-905, 2000.

[17] H. Tao, H. Sawhney and R. Kumar, "Object Tracking with Bayesian Estimation of Dynamic Layer Representations," IEEE Trans. PAMI, 24(1): 75-89, 2002.

[18] J. Wang and E. Adelson, "Layered Representation for Motion Analysis," In CVPR, p361-366, 2003.

[19] J. Yedidia, W. Freeman and Y. Weiss, "Generalized belief propagation," In NIPS, 2000.

[20] Z. Yin and R. Collins, "Moving Object Localization in Thermal Imagery by Forward-backward MHI," In CVPR workshop on Object Tracking and Classification in and Beyond the Visible Spectrum (OTCBVS), 2006.