Ensemble Video Object Cut in Highly Dynamic Scenes

Xiaobo Ren, Tony X. Han, and Zhihai He
Department of Electrical and Computer Engineering
University of Missouri, Columbia MO 65211
xr7rf@mail.missouri.edu {hantx, hezhi}@missouri.edu

Abstract

We consider video object cut as an ensemble of frame-level background-foreground object classifiers which fuses information across frames and refine their segmentation results in a collaborative and iterative manner. Our approach addresses the challenging issues of modeling of background with dynamic textures and segmentation of foreground objects from cluttered scenes. We construct patch-level bag-of-words background models to effectively capture the background motion and texture dynamics. We propose a foreground salience graph (FSG) to characterize the similarity of an image patch to the bag-of-words background models in the temporal domain and to neighboring image patches in the spatial domain. We incorporate this similarity information into a graph-cut energy minimization framework for foreground object segmentation. The background-foreground classification results at neighboring frames are fused together to construct a foreground probability map to update the graph weights. The resulting object shapes at neighboring frames are also used as constraints to guide the energy minimization process during graph cut. Our extensive experimental results and performance comparisons over a diverse set of challenging videos with dynamic scenes, including the new Change Detection Challenge Dataset, demonstrate that the proposed ensemble video object cut method outperforms various state-of-the-art algorithms.

1. Introduction

Detecting and segmenting moving objects from the background is the enabling step in intelligent video analysis [23, 11]. A number of methods and algorithms have been developed for background subtraction and moving object detection [15, 23]. However, accurate and reliable moving object detection from cluttered and highly dynamic background remains as a challenging problem. Existing work has been focusing on and mainly evaluated with indoor and outdoor facility surveillance videos which often have relatively stable scenes. However, videos captured in natural environments represent a large class of challenging scenes that have not been sufficiently addressed in the literature [11]. These types of scenes are usually cluttered and dynamic with swaying trees, ripping water, moving shadows and sun spots, rain, etc. The key challenge here is how to establish effective models to capture the complex background motion and texture dynamics.

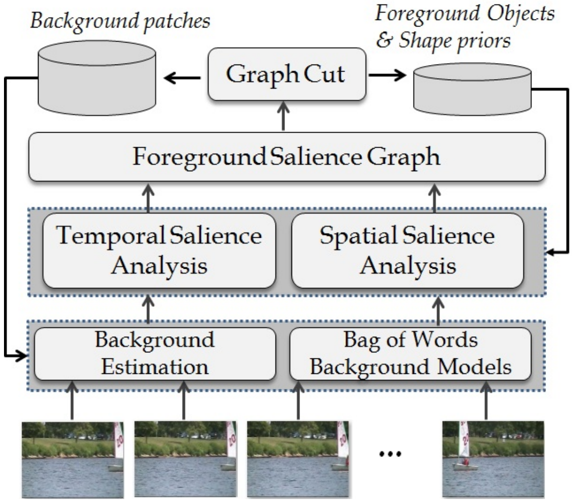
Background patches
Foreground Objects
& Shape priors
Graph Cut
Foreground Salience Graph
Temporal Salience Analysis
Spatial Salience Analysis
Background Estimation
Bag of Words Background Models
...

Figure 1: Overview of the object detection system using ensemble video object cut.

In this work, we consider video object segmentation as an ensemble of frame-level background-foreground object classifiers which fuses information across frames and refine their segmentation results in a collaborative and iterative manner, as illustrated in Fig. 1. Our approach integrates patch-level local background modeling with bags of words, region-level foreground object segmentation with graph cuts, and temporal domain information fusion among foreground-background classifiers at neighboring frames. Our extensive experimental results and performance comparisons over a diverse set of challenging videos, including the recent Change Detection Challenge Dataset [7], demonstrate that the proposed method outperforms various state-

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.254

IEEE computer society

of-the-art algorithms.

2. Bag of Words Models for Dynamic Backgrounds

The remainder of the paper is organized as follows. Section 3 reviews the related background segmentation work. Section 4 provides an overview of the proposed system. Sections 2 and 5 present the bag-of-words background models, foreground salience map, and our graph-cut algorithm for foreground object segmentation. The experimental results are presented in Section 7. Concluding remarks and discussions are given in Section 8.

3. Related Work

There is a significant body of research conducted during the past two decades on background modeling and foreground object detection. However, the availability of methods that are robust and generic enough to handle the complexities of most natural dynamic scenes is still very limited [15, 16]. Early work on background subtraction operated on the assumption of stationary background [26]. To handle motion in the background, methods with pixel-level motion matching and background model relaxation within the pixel neighborhood have been investigated. For example, a non-parametric technique was proposed in [4] for estimating background probabilities using Kernel density functions. This method addressed the issue of nominally moving cameras using a local search for the best match for each incident pixel in neighboring models. Ren *et al.* explicitly addressed the issue of background subtraction in a non-stationary scene by introducing the concept of a spatial distribution of Gaussians (SDG) [20]. In [11], distributional signatures and local warping methods have been studied. In [10], for each pixel, it builds a codebook consisting of one or more codewords. Samples at each pixel are clustered into the set of codewords based on a color distortion metric together with brightness bounds.

Our work is also related to graph-cut based image segmentation. Considering spatial context and neighborhood constraints, graph cut optimization has achieved fairly good performance in image segmentation [2]. Iterated graph cut is used in [21] to search over a nonlocal parameter space. Background cut is proposed in [24] which combines background subtraction and color/contrast based models.

We recognize that, for accurate and robust video object detection and segmentation in dynamic scenes, background modeling of the dynamic pixel process at the image patch level, spatial context analysis and graph cut optimization at the region-level, and ensemble foreground-background classification at the sequence level should be jointly considered. In this work, we propose to establish a new framework which tightly integrates these three important components for accurate and robust video object cut in highly dynamic scenes.

4. Overview of the Proposed Approach

The basic flow of the proposed ensemble video object cut method is shown in Fig. 1. We first scan the image sequence, perform initial background-foreground image patch classification, and construct bag-of-words (BoW) background models with Histogram of Oriented Gradients (HOG) features. This BoW model is able to capture the background motion and texture dynamics at each image location. To segment the foreground object, for each image patch, we develop features to describe its texture and neighborhood image characteristics. Based on the BoW background models, we analyze its temporal salience. We also compare the image patch to its neighborhood patches to form the spatial salience measure. Based on this spatiotemporal salience analysis results, we construct the foreground salience graph. We then apply the graph-cut energy minimization method to obtain the foreground segmentation. These background-foreground classification results of neighboring frames are fused together to further update the weights of the foreground salience graph. Shape prior information is extracted from the detected foreground objects and used as constraints to guide the graph-cut energy minimization procedure. This classification-fusion-refinement procedure is performed in an iterative manner to achieve the final video object segmentation results.

Most of the existing background models are constructed at the pixel level [4, 26, 10, 11]. These methods typically assume that the pixel processes at different pixel locations are independent of each other. We recognize that, without considering image characteristics in the pixel neighborhood, this type of methods often produce inconsistent decision and are not resilient to image noise and background motion.

In this work, we propose to develop background models at the patch level using BoW features. Let $\{F_1, F_2, \cdots, F_N\}$ be the sequence of images to be analyzied. Let $P_n^{(x,y)}$ be the patch extracted from frame F_n at location (x, y). We extract its HOG feature, denoted by $f_n^{(x,y)}$, from the image patch $P_n^{(x,y)}$. We choose the HOG feature because it is able to effectively capture the texture information of image patches and is relatively invariant under changes of lighting conditions. We observe that the set of co-located image patches $\{P_n^{(x,y)} | 1 \leq n \leq N\}$ will have a complicated correlation structure in the high-dimensional feature space. To capture this correlation structure, we use the complete-linkage clustering algorithm [5] to cluster image patch features $\{f_n^{(x,y)}\}$ and obtain the visual words. Using these visual words, we can then construct a histogram, denoted by $h_n^{(x,y)}$ to describe the image patch $P_n^{(x,y)}$.

5. Foreground Salience Graph and Graph Cut

In foreground object detection, we need to detect those image patches which are salient in comparison with background models on both appearance and texture dynamics. In this work, we propose to construct a foreground salience graph (FSG) to characterize the salience of an image patch in the spatiotemporal domain. We will then formulate the object segmentation as an energy minimization problem which can be solved using the graph cut method.

5.1. Foreground Salience Graph

The FSG consists of two components: temporal salience and spatial salience. During the design of these two components, we aim to find a balance between global smoothness and local details in video object segmentation.

Temporal salience map. The temporal salience measures the dis-similarity between the current image patch $P^{(x,y)}$ and the background model. Let $d(P^{(x,y)}, P_k^{(x,y)})$ be the feature distance between the current image patch and co-located background image patch $P_k^{(x,y)}$ at frame k. We adopt the χ^2-distance

$$d(P^{(x,y)}, P_k^{(x,y)}) = \frac{1}{2}\sum_i \frac{(g(i)-h(i))^2}{g(i)+h(i)}, \qquad (1)$$

where g and h are the BoW histogram features describing image patch $P^{(x,y)}$ and its co-located background image patch $W_k^{(x,y)}$, respectively. The temporal salience at location (x,y) is the defined as

$$D^t(P^{(x,y)}) = D^t(x,y) = \min_k d(P^{(x,y)}, P_k^{(x,y)}). \qquad (2)$$

Fig. 2 shows temporal salience maps for four example frames from the *Camera Trap* dataset. The camera data contains very challenging videos with highly dynamic scenes with large tree waving motion, strong moving shadow and sunlight spots. A detailed description of this dataset is provided in Section 7. Here, red and blues pixels represent image patches with large and small temporal salience values, respectively. We can see that the bag-of-words background model and the temporal salience map are able to efficiently characterize the complex background motion

Spatial salience map. As discussed in Section 3, to achieve consistently accurate and reliable foreground object segmentation, we need to consider the spatial context of the image patch and the image characteristics in its spatial neighborhood. This is particularly important in highly cluttered and dynamic scenes. To form the spatial salience measure between two neighboring image patches, $P^{(x,y)}$ and $Q^{(x',y')}$, we analyze both color and texture information.

As illustrated in Fig. 3, at the middle point O, we consider a circular neighborhood, which is partitioned into two half-discs with an orientation angle θ [1]. In our experiments, the radius of the circle is set to be the same as the patch width. The partition line (or the angle θ) should be perpendicular to the line connecting these two patch centers (x,y) and (x',y'). We notice that the patch boundary may not align well with the object boundary. We allow the middle point O, as well as the partition angle θ to vary within a small neighborhood. We then find the χ^2-distance between the color histograms of these two half-discs, denoted by $D^s[O,\theta](P,Q)$. The spatial salience measure between patches $P^{(x,y)}$ and $Q^{(x',y')}$ is then defined as

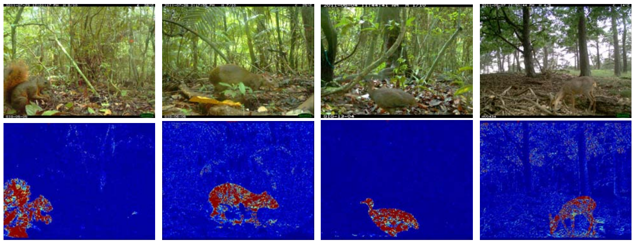

Figure 2: the first row: original sample images from *Camera_Trap* dataset; the second row: temporal salience maps with red and blue pixels representing large and small temporal salience values, respectively.

$$D^s(P^{(x,y)}, Q^{(x',y')}) = \max_O \max_\theta D^s[O,\theta](P,Q). \qquad (3)$$

To effectively differentiate the background and foreground textures, we propose to modulate this spatial salience with LBP texture weights. More specifically, as illustrated in Fig. 3, at image location (x,y), we construct its LBP descriptor by comparing its average intensity against its eight neighbors. We denote the LBP descriptor of the current image as $LBP(x,y)$ and those of the k-th background model as $LBP_k(x,y)$. We define the LBP texture weight as

$$w_H(x,y) = \min_k d_H[LBP(x,y), LBP_k(x,y)], \qquad (4)$$

where $d_H[\cdot,\cdot]$ represents the Hamming distance between two LBP binary vectors. This LBP texture weight aims to find a balance between effectively differentiating the foreground and background image textures and accomondating background motions.

Foreground Salience Graph. Based on the temporal and spatial salience measures, we can then construct the foreground salience graph. We represent the image by an 8-connectivity undirected graph $G(V,\mathcal{E})$, where V is the set of all image patches in the current frame F_N. $\mathcal{E}$ representing all the adjacent or 8-connected pairs of nodes in G. This type of links are called N-links [2]. In addition, for background-foreground segmentation purposes, we also introduce two terminal nodes, the foreground t and background nodes s. As illustrated in Fig. 3(b), all nodes in V are connected to these two terminal nodes. The cor-

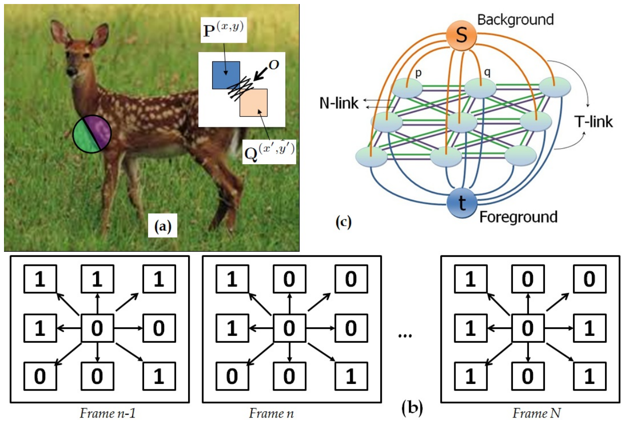

Figure 3: (a) Given an intensity image, at a given location, we can find the optimum angle θ which maximize the distance between the histograms of oriented gradients in these two half discs; (b) LBP texture weights; (c) foreground salience graph.

responding links are referred as T-links. The segmentation of the image is represented by a set of binary labels $X = \{x_p | x_p \in \{0, 1\}\}$, where p represents a node in the graph or an image patch.

We formulate the foreground object segmentation as a graph-cut energy minimization problem, which aims to minimize the following global energy function:

$$\begin{aligned} E(X) &= \sum_{p \in V} E_p(x_p) \cdot w_H(x_p) + \lambda_1 \sum_{(p,q) \in \mathcal{E}} E_D(x_p, x_q) \\ &+ \lambda_2 \sum_{(p,q) \in \mathcal{E}} E_S(x_p, x_q). \end{aligned} \tag{5}$$

Here, $E_p(x_p)$ represents the T-link energy while $E_D(x_p, x_q)$ and $E_S(x_p, x_q)$ represent the N-link energy. λ_1 and λ_2 are constant parameters that balance the influence of these three energy terms. $w_H(x_p)$ is the LBP texture weight defined in (4). The T-link energy provides an initial assessment if the image patch belongs to the background or foreground, which is defined based on the temporal salience measure:

$$E_p(x_p) = \begin{cases} \frac{D^t(p)}{\alpha_1}, & x_p = 0; \\ \alpha_2 (1 - \frac{D^t(p)}{20\alpha_1}), & x_p = 1; \end{cases} \tag{6}$$

where α_1 and α_2 are constants. The N-link energy represents comparison between neighboring patches. $E_D(x_p, x_q)$ captures the discontinuity between two neighboring patches in the temporal salience map, which is defined as

$$E_D(x_p, x_q) = \gamma \cdot e^{-\beta_D \cdot |D^t(p) - D^t(q)|}, \tag{7}$$

where β_D is a normalization term

$$\beta_D = [\sum_{p \in V} D^t(p)]^{-1}. \tag{8}$$

Another component of the N-link energy, $E_S(x_p, x_q)$, is defined based on the spatial salience,

$$E_s(x_p, x_q) = \gamma \cdot e^{-\beta_S \cdot D^s(p,q)}, \tag{9}$$

where β_S is a normalization term computed as

$$\beta_S = [\max_{(p,q) \in \mathcal{E}} D^s(p, q)]^{-1}. \tag{10}$$

In this work, we use the min-cut method [2] to minimize the global energy function in (5). The output of this graph-cut minimization procedure will be the foreground object segmentation.

6. Iterative Ensemable Video Object Cut

We recognize that, in cluttered scenes, the initial segmentation often yields incorrect segmentation and object contours. For example, in the Camera-trap dataset, we find that some parts of the animal body are well segmented in some video frames but poorly segmented in other frames since the foreground object has moved to different background regions. Motivated by this observation, we propose to consider the problem of video object cut as an ensemble of frame-level foreground-background classifiers, which share and fuse the classification information across frames, helping each other to refine the segmentation in an iterative manner. To this end, we will explore two major ideas.

6.1. Foreground Probability Map

From the existing foreground-background classification results of all frames, we estimate the foreground-background probability map for each frame. More specifically, with the new background masks, we remove those false background image patches from the background model. For each image patch P at location x_P, we update its minimum distance $D^t(P)$ to all background image patches in the model using (2). With the new foreground masks, we can also construct bag-of-words models for the foreground objects. Following the procedure in Section 2, we can then define the foreground temporal salience measure $D^t_f(P)$ for a given image patch P in the current frame, which will measure the similarity between the current patch and detected foreground patches. We define the foreground probability map as

$$\gamma(P) = 1 - e^{-D_t(P)/D^t_f(P)}, \tag{11}$$

which measures the probability of P to be a foreground patch. We then use $\gamma(P)$ as a weighting factor to update the weights of t-link (edges to the foreground) and the N-link edges as follows:

$$\begin{aligned} w^{i+1}_{t-link}(P) &= w^{i}_{t-link}(P)\cdot[1+c_0\gamma(P)], \quad (12)\\ w^{i+1}_{n-link}(P,Q) &= w^{i}_{t-link}(P)\cdot(1+c_1|\gamma(P)-\gamma(Q)|). \end{aligned}$$

6.2. Foreground Shape Priors

The graph cut (or s/t cut) problem can be solved by finding a maximum flow from the source s (background) to the destination t (foreground). The theorem of Ford and Fulkerson [6] states that a maximum flow from s to t saturates a set of edges in the graph dividing the nodes into two disjoint parts $\{\mathcal{S},\mathcal{T}\}$ corresponding to a minimum cut. One typical approach to solve the max-flow problem is to use augmenting paths [6], for which fast algorithms are available [4] with linear running time. The algorithm starts with zero flow and gradually increases the flow amount as long as there is an open path from the source to the destination on the residual graph. The incremental equals to the minimum of the residual capacities on the path, as illustrated in Fig. 4(b). The set of saturation edges corresponds to the final graph cut result [6]. In our case, the object contour obtained from graph cut segmentation will run across these saturation edges. Therefore, an incorrect selection of the saturation edge in the final stage will result in incorrect segmentation, causing distorted object shapes or contours. Fig. 4(a) shows one example of incorrect segmentation, where the leg of the deer is missing in frame while it is well segmented in another frame. To address this issue, we propose to use the object segmentation results from other frames, extract shape prior information of the foreground object, and use this information to guide the graph cut algorithm. More specifically, to construct the shape prior, we denote the shape segment with the image patch P by $S[P]$. Let $\theta(S[P])$ be its orientation angle quantized to 8 bins between 0 and 2π. Let

$$\mathcal{N}[P]=\{Q_1,\cdots,Q_K\} \quad (13)$$

be the K nearest neighbor image patches of P. From the existing segmentation results, we estimate the following conditional probability

$$p(\theta)=p\{\theta(S[P])\,|\,\theta(S[Q_1]),\cdots\theta(S[Q_K])\}. \quad (14)$$

This probability $p(\theta)$ attempts to predict the shape segment orientation based on the foreground object shapes segmented from other frames. Using this probability, we are able to determine the most likely position of the saturation edge in the augmenting path. If the residual capacity of this edge in the residual graph is below a certain threshold, we can impose early termination of the augmenting process and set this edge as the saturation edge.

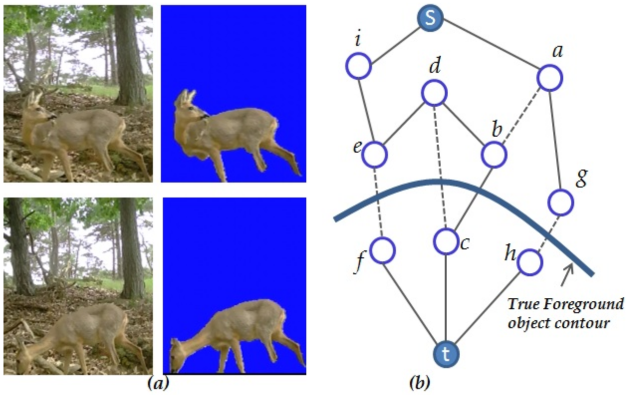

Figure 4: shape prior assisted graph cut.

7. Experimental Results

We evaluate our method with a diverse set of publicly available challenging videos used in the literature and compare our method with the state-of-the-art methods.

7.1. Datasets

The data used for our performance evaluation consists of existing benchmark datasets and our in-house wildlife monitoring videos: (A) **WavingTrees** [25]; (B) **Fountain** [23]; (C) **Combination** with 9 complex scenes [13, 12]; (D) **WaterObject** with a jug floating on the rippling water [27]; (E) **RainCar** with several cars passing through a heavily raining scene [16]; and (G) the new Change Detection Challenge dataset [7] for performance evaluations and comparisons of moving object detection and segmentation algorithms, where the most recent results are available. We will also evaluate our method with (F) the **Camera_Trap** dataset of wildlife monitoring videos. In our on-going work on automated large-scale wildlife monitoring, we have collected over 1 million camera-trap images of wildlife species. This dataset consists of 23 species of wildlife animals captured by camera-traps, in both daytime color and nighttime infrared formats. These are very challenging videos with highly cluttered and dynamic wooded scenes. This dataset will be made available online for public use.

In our experiments, we use a patch size of 32×32. The number of patches used for background modeling ranges from 2560 to 7680 depending on the video size. The dictionary size is 128. We use a group of 10-15 video frames as a segmentation unit for ensemble video object cut.

7.2. Quantitative Evaluations

In this section, we provide quantitative evaluations measured using the F-score:

$$F=\frac{2\times recall\times precision}{recall+precision}=\frac{2TP}{2TP+FN+FP},$$

where TP stands for true positive, FP stands for false positive, FN stands for false negative [13]. Table 1 shows the F-score results on the **Combination** dataset in comparison with (a) **MM03**, the Bayes model approach by Li *et al.* [12]; (b) **LBP-P**: the pixelwise LBP histogram based approach by Heikkil *et al.* [8]; and (c) **PKDE**: the pattern kernel density estimation method by Liao [13] with different parameter settings. The best performance score is highlighted in bold. It can be seen that the proposed method outperforms other methods in the literature on most test sequences. For sequences *Escalator, Fountain, and Shopping Mall*, our F-scores are just slightly lower than the PKDE method [13].

Next, we provide comprehensive performance evaluations on the Dynamic Background videos from the Change Detection Challenge dataset [7]. The website also publishes results on this dataset by an extensive list of methods recently developed in the literature. In this work, we include the top 7 methods for performance comparison: (a) **CP**, the Chebyshev probability approach [17]; (b) **Feedback**, the feedback approach [9]; (c) the **ViBe+** method [3]; (d) **PSP-MRF**, the probabilistic superpixel Markov random fields approach [22]; (e) **KDE**, the integrated spatiotemporal approach [18]; (f) **QCH**, the quasi-continuous histogram approach [19]; and (g) **SC-SOBS**, the SOBS algorithm [14]. The definition of Specificity, FPR (False Positive Rate), FNR (False Negative Rate), PWC (Percentage of Wrong Classifications), F-Score, and Precision are provided on the website [7]. From Table 2, we can see that our method significantly outperforms other methods. For example, we have achieved an average precision of 95.34%, much higher than the second best 83.26% [9]. This is because our method is able to effectively capture and model the highly dynamic background motion and to accurately locate the object boundary by sharing and fusing foreground-background classification information between frames.

7.3. Qualitative Evaluations

In this section, we provide qualitative evaluations of our ensemble video object cut method and performance comparisons with other state-of-the-art methods in the literature. Fig. 5 shows one example of segmentation results by our method on the *Fountain* dataset in comparison with the Bayesian modeling approach [23]. We can see that both the Bayesian modeling approach and our approach are able to accurately model the background water motion. However, their method tends to under-segment the foreground person with the low-contrast waist and hair areas being classified as background. Our method is able to accurately detect and segment the whole person with graph-cuts by considering the spatial context. Fig. 6 shows the results on the **Combination** dataset in comparison with the ACMMM03 [12] and LBP-P [8] methods. We can see that the proposed method yields more accurate and robust foreground object detection and segmentation than these two methods. Fig. 7 shows how our ensemble video cut method is able to refine the segmentation results in an iterative manner by sharing and fusing foreground-background classification information between frames.

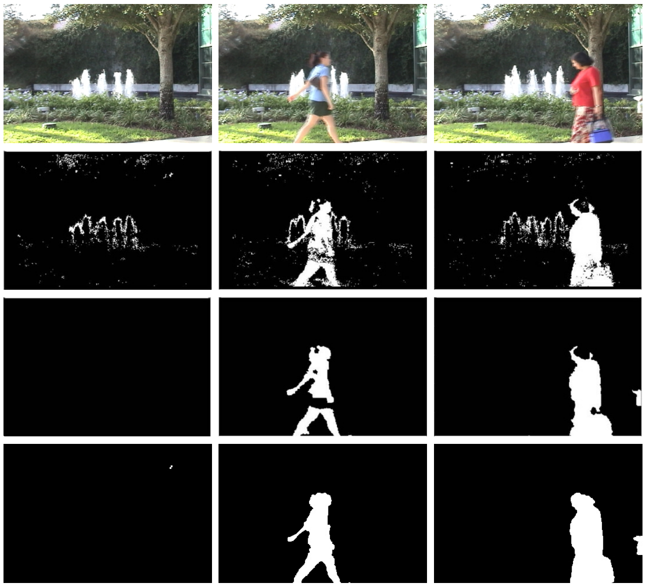

Figure 5: The top row are the original images from *Fountain* dataset [23]. The second row are the results obtained by using Bayesian modeling [23]. The third row are results obtained by the proposed method.

8. Conclusion

In this work, we have successfully developed a video object segmentation scheme for highly dynamic and cluttered scenes. Our approach integrates patch-level local background modeling with bags of words, region-level foreground object segmentation with graph cuts, and temporal domain information fusion among foreground-background classifiers at neighboring frames. We constructed patch-level bag-of-words background models to effectively capture the background motion and texture dynamics. We have developed a foreground salience graph (FSG) to characterize the similarity of an image patch to the bag-of-words background models in the temporal domain and to neighboring image patches in the spatial domain. We incorporated this similarity information into a graph-cut energy minimization framework for foreground object segmentation. The major novelty of this paper lies in considering the video object segmentation in challenging natural scenes as an ensemble of frame-level background-foreground object classifiers, which fuses information across frames and re-

Table 1: Performance comparison with the F-score (%) on the **Combination** dataset with other methods.

Sequences	MM03	LBP-P	PKDE$_{ltp}$	PKDE$_{siltp}$	PKDE$^{w=2}_{mb-siltp}$	PKDE$^{w=3}_{mb-siltp}$	PKDE$^{w=1+2+3}_{mb-siltp}$	This Work
AirportHall	50.18	50.29	62.13	68.14	65.87	63.60	68.02	**81.65**
Bootstrap	60.46	52.00	73.86	75.35	69.45	64.87	72.90	**76.14**
Curtain	56.08	71.42	74.19	91.16	89.37	87.97	92.40	**93.63**
Escalator	32.95	53.93	67.71	63.90	64.37	60.18	**68.66**	68.24
Fountain	56.49	75.33	81.05	83.45	81.17	77.60	**85.04**	84.00
ShoppingMall	67.84	62.92	73.91	79.62	77.75	74.49	**79.65**	79.46
Lobby	20.35	52.34	77.85	78.80	73.82	67.16	79.21	**83.86**
Trees	75.40	60.57	42.98	42.54	51.88	61.53	67.83	**89.12**
WaterSurface	63.66	82.21	41.46	74.30	81.08	83.51	83.15	**94.95**
Total	59.21	63.46	67.59	75.35	75.24	73.59	78.69	**83.45**

Table 2: Performance comparison on Dynamic Background videos in the Change Detection dataset with other methods.

Method	Avg Recall	Avg Specificity	Avg FPR	Avg FNR	Avg PWC	Avg F-Score	Avg Precision
CP [17]	0.8182	0.9982	0.0018	0.1818	0.3436	0.7656	0.7633
Feedback [9]	0.6955	**0.9989**	**0.0011**	0.3045	0.5394	0.6829	0.8326
ViBe+ [3]	0.7616	0.9980	0.0020	0.2384	0.3838	0.7197	0.7291
PSP-MRF [22]	0.8955	0.9859	0.0141	0.1045	1.4514	0.6960	0.6576
KDE [18]	0.8401	0.9908	0.0092	0.1599	1.1501	0.6016	0.5413
QCH [19]	0.8909	0.9896	0.0104	0.1091	1.1301	0.6430	0.5347
SC-SOBS [14]	0.8918	0.9836	0.0164	0.1082	1.6899	0.6686	0.6283
This Work	**0.9470**	0.9978	0.0022	**0.0530**	**0.3255**	**0.9496**	**0.9534**

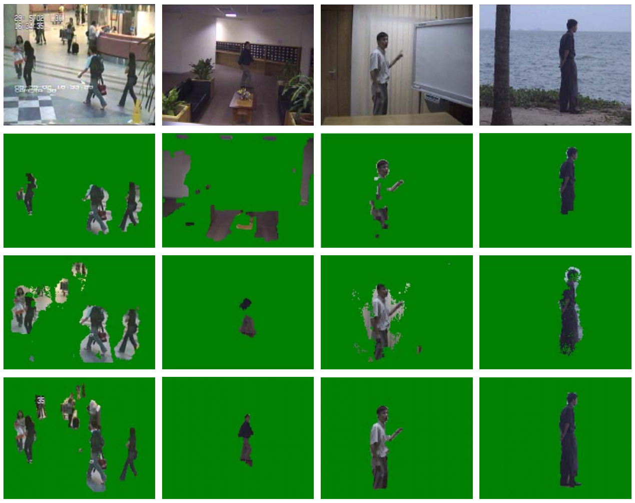

Figure 6: Segmentation results on four videos from the **Combination** dataset. First row: original image from dataset *AirportHall*, *Lobby*, *Curtain* and *WaterSurface* (from left to right); Second row: ACMMM03 [12]; Third row: LBP-P [8]; Last row: proposed MBGC

fine their segmentation results in a collaborative and iterative manner, achieving surprising good performance. Our extensive experimental results and performance comparisons over a diverse set of challenging videos with dynamic scenes, including the Change Detection Challenge Dataset, demonstrated that the proposed ensemble video object cut method outperforms various state-of-the-art algorithms.

9. Acknowledgements

This work was partly supported by NSF under the grants DBI-1062354 and EF-1065749.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5), 2011.

[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut maxflow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

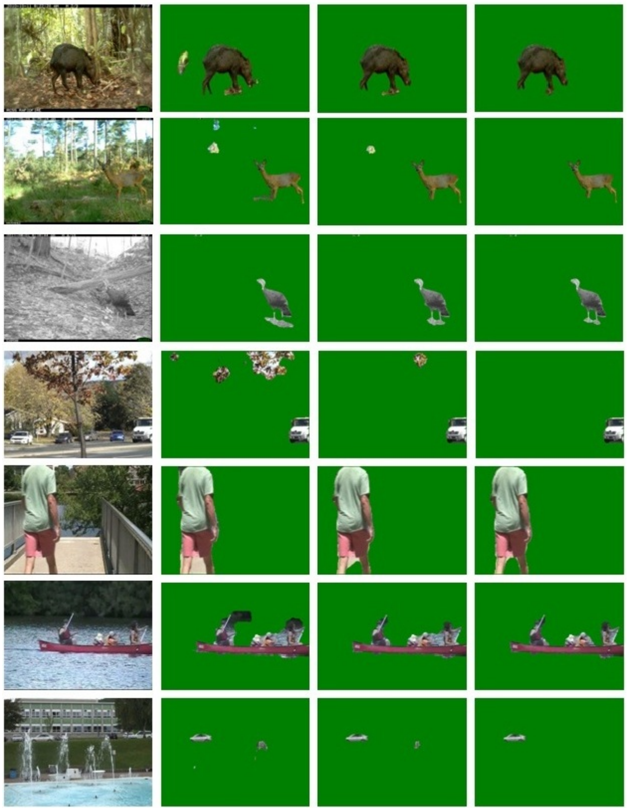

Figure 7: Iterative ensemble video obejct cut; the first row: the original video frames; the second row, the segmentation results after the first iteration; the third row, the segmentation results after the first iteration; the fourth row, the final segmentation results.

[3] M. V. Droogenbroeck and O. Paquot. Background subtraction: Experiments and improvements for vibe. *IEEE Workshop on Change Detection, CVPR*, 2012.

[4] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. *ICCV*, 2000.

[5] B. S. Everitt, S. Landau, and M. Leese. Cluster analysis (fourth ed.). *London: Arnold*, 2001.

[6] L. Ford and D. Fulkerson. Flows in networks. *Princeton University Press*, 1962.

[7] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. changedetection.net: A new change detection benchmark dataset. *Proc. IEEE Workshop on Change Detection (CDW12) at CVPR12*, 2012.

[8] M. Heikkil, M. Pietikinen, and S. Member. A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. Pattern Anal. Machine Intell*, 28, 2006.

[9] M. Hofmann, P.Tiefenbacher, and G. Rigoll. Background segmentation with feedback: The pixel-based adaptive segmenter. *IEEE Workshop on Change Detection, CVPR*, 2012.

[10] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground-background segmentation using codebook model. *Real-time Imaging*, 11(3):167–256, 2005.

[11] T. Ko, S. Soatto, and D. Estrin. Background subtraction with distributions. *In Proceedings of the European Conference on Computer Vision*, 2008.

[12] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian. Foreground object detection from videos containing complex background. *Proceedings of the eleventh ACM international conference on Multimedia*, pages 2–10, 2003.

[13] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. *ICCV*, pages 1301–1306, 2010.

[14] L. Maddalena and A. Petrosino. The sobs algorithm: what are the limits? *IEEE Workshop on Change Detection, CVPR*, 2012.

[15] V. Mahadevan and N. Vasconcelos. Background subtraction in highly dynamic scenes. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2008.

[16] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):171–177, 2010.

[17] A. Morde, X. Ma, and S. Guler. Learning a background model for change detection. *IEEE Workshop on Change Detection, CVPR*, 2012.

[18] Y. Nonaka, A. Shimada, H.Nagahara, and R. Taniguchi. Evaluation report of integrated background modeling based on spatio-temporal features. *IEEE Workshop on Change Detection, CVPR*, 2012.

[19] O.Strauss, D. Sidib, and W. Puech. Quasi-continuous histogram based motion detection. *Technical Report, LE2I*, 2012.

[20] Y. Ren, C.-S. Chua, and Y.-K. Ho. Motion detection with nonstationary background. *Machine Vision and Application, Springer-Verlag*, 2003.

[21] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23:309–314, 2004.

[22] A. Schick, M.Bum, and R.Stiefelhagen. Improving foreground segmentations with probabilistic superpixel markov random fields. *IEEE Workshop on Change Detection, CVPR*, 2012.

[23] Y. Sheikh and M. Shah. Bayesian modeling of dynamic scenes for object detection. *IEEE PAMI*, 27:1778–1792, 2005.

[24] J. Sun, W. Zhang, X. Tang, and H. Y. Shum. Background cut. *In European Conference on Computer Vision*, pages 628–641, 2006.

[25] K. Toyama, J. Krumm, B. Brumitt, and M. Brian. Wallflower:principles and practice of background maintenance. *In European Conference on Computer Vision*, 255(1), 1999.

[26] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

[27] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust kalman filter. *ICCV*, 2003.