Saliency-Aware Geodesic Video Object Segmentation

Wenguan Wang [1], Jianbing Shen* [1], Fatih Porikli [2]

[1]Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, China
[2]Research School of Engineering, Australian National University, and NICTA Australia

Abstract

We introduce an unsupervised, geodesic distance based, salient video object segmentation method. Unlike traditional methods, our method incorporates saliency as prior for object via the computation of robust geodesic measurement. We consider two discriminative visual features: spatial edges and temporal motion boundaries as indicators of foreground object locations. We first generate frame-wise spatiotemporal saliency maps using geodesic distance from these indicators. Building on the observation that foreground areas are surrounded by the regions with high spatiotemporal edge values, geodesic distance provides an initial estimation for foreground and background. Then, high-quality saliency results are produced via the geodesic distances to background regions in the subsequent frames. Through the resulting saliency maps, we build global appearance models for foreground and background. By imposing motion continuity, we establish a dynamic location model for each frame. Finally, the spatiotemporal saliency maps, appearance models and dynamic location models are combined into an energy minimization framework to attain both spatially and temporally coherent object segmentation. Extensive quantitative and qualitative experiments on benchmark video dataset demonstrate the superiority of the proposed method over the state-of-the-art algorithms.

1. Introduction

Unsupervised video object segmentation methods aim at automatically extracting the object from the whole video. Such segmentation has shown to benefit many specific visual tasks and applications, such as video summarization, compression and human-computer interaction to name a few. Appearance information and motion cues are usually employed by video segmentation approaches. Some works in [6, 17, 12] analyzed point trajectories in order to take advantage of motion information available in multiple frames. Brox *et al.* [6] offered a framework for trajectory-based video segmentation through building affinity matrix between pairs of trajectories. Lezama *et al.* [17] grouped pixels with coherent motion computed via long-range motion vectors from the past and future frames. Another approach by Fragkiadaki *et al.* [12] detected discontinuities of embedding density between spatial-neighboring trajectories. As the work [15] pointed out, these trajectory-based techniques suffer from the challenges associated with tracking (drift, occlusion and initialization) and clustering (model selection and computational complexity) and lack of prior information for a successful object segmentation. Some efforts [5, 26, 30] presented efficient optimization frameworks for bottom-up final segmentation employing both appearance and motion cues.

Recently, several methods [15, 19, 32] explored the notion of what a foreground object should look like in video data. These approaches generate considerable object proposals [11, 8] in every frame and transform the task of video object segmentation into an object region selection problem. In this selection process, both motion and appearance information are combined to measure the *objectness* of a proposal. More specifically, a clustering process was introduced for finding objects by Lee *et al.* [15], a constrained maximum weight cliques technique to model the selection process was proposed by Ma and Latecki [19], and a layered directed acyclic graph based framework was presented by Zhang *et al.* [32]. However, these proposal based techniques have high computational complexity, and their dependency on the large number of proposals leads to much difficulty and complexity of the selection process.

Our goal is to segment the foreground objects from the

*Corresponding author: Jianbing Shen (shenjianbing@bit.edu.cn). This work was supported in part by the National Basic Research Program of China (973 Program) (No. 2013CB328805), the National Natural Science Foundation of China (No. 61272359), the Australian Research Council's Discovery Projects funding scheme (project DP150104645), and the Program for New Century Excellent Talents in University (NCET-11-0789). Specialized Fund for Joint Building Program of Beijing Municipal Education Commission.

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

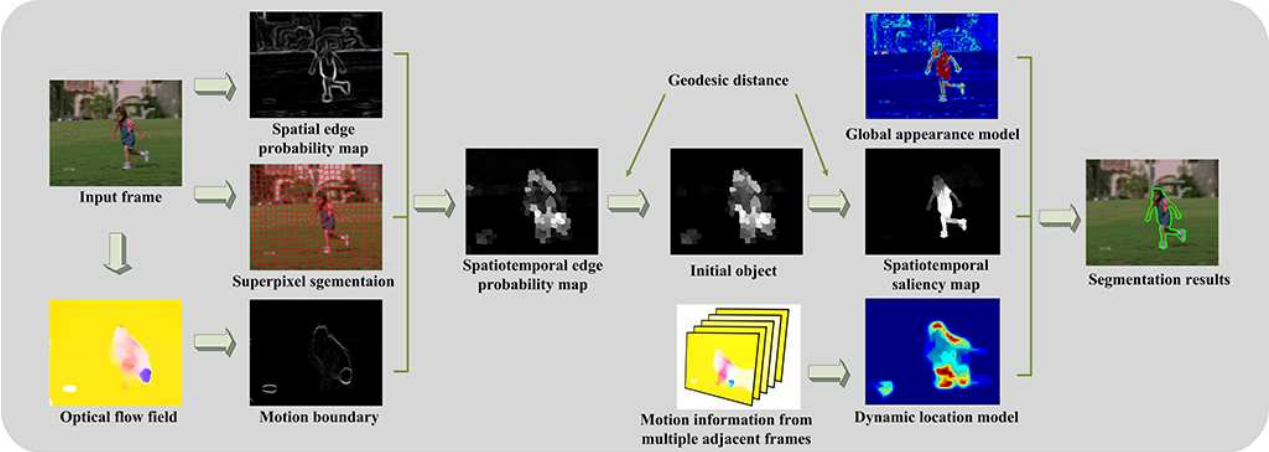

Figure 1. Overview of saliency-aware geodesic video object segmentation.

background in all frames of a given video sequence without any user annotation and semantic prior. Our method is based on the proposed visual saliency detection technique that incorporates several visual cues such as motion boundary, edge and color. Object and background estimations generated by our method provide consistent and reliable priors for higher level object segmentation tasks. This topic is less explored, mainly due to only a few methods specifically designed for video saliency till now. These saliency methods [14, 20, 28, 26, 13, 21], however, usually build their system as a simple combination of existing image saliency models with motion cues. Furthermore, the performance of these methods is not good enough to guide the segmentation. Our method correctly estimates the locations of object and background and gains uniform saliency maps. On the other hand, our video object segmentation algorithm is based on the geodesic distance, which has been proved to be effective for interactive image and video segmentation with user brushes [3, 25, 2, 10]. However, in many vision applications, such as processing a large number of video data, it is usually tedious and impractical for users to handle the video frames manually. In this paper, we try to introduce geodesic distance into our totally automatic segmentation framework, which is different with previous approaches [3, 25, 2, 10] that require careful user assistance.

2. Our approach

Fig. 1 shows an overview of our approach. First, input frames are oversegmented into superpixels. For each superpixel, two types of edges are extracted: spatial static edges within the same frame and motion boundary edges estimated from neighboring frames. Geodesic distance, which is defined as the shortest paths between two superpixels on the image, is then adopted in a intra-frame graph for computing the object probability of each superpixel. Based on the observation that the object areas are surrounded by the regions with high spatiotemporal edge value, the object probability is computed as the shortest geodesic distance to the frame boundaries. A self-adaptive threshold is used to obtain initial labeling of the frame into background and foreground regions. Next an inter-frame graph is constructed for producing spatiotemporal saliency maps by the computation of geodesic distance to the estimated background regions of two adjacent frames. Finally, to achieve refined estimation of foreground, global appearance model for foreground and background is established by saliency results. Dynamic location model for each frame is estimated from motion information extracted from few subsequent frames. Spatiotemporal saliency maps, global appearance model and dynamic location model are combined into an energy function for final segmentation. Our source code will be publicly available online [1].

2.1. Object estimation using spatiotemporal edges

Edges provide good guide in predicting object boundaries, while simultaneously being very efficient. Motion information also offers a simplified but very effective indicator of object, the pixels which change abruptly from neighbors often gain more attention. As shown in Fig. 1, the location of static edges for single frame and the optical flow field estimated from two consecutive frames could provide useful information for detecting object. We base our approach on these two discriminative features for priming object locations.

Given an input video sequence $\mathbf{F} = \{F^1, F^2, \cdots\}$, we compute an edge probability map $E_c^k(x_i^k)$ corresponding to k-th frame F^k at pixel x_i^k using [16]. The optical flow between pairs of subsequent frames are obtained by the large displacement motion estimation algorithm [7]. Let V^k be the optical flow field of frame F^k, we then compute the gradient magnitude E_o^k of the optical flow field V^k as

[1] http://github.com/shenjianbing/videoseg15

$E_o^k = \|\nabla V^k\|$. We oversegment each frame into superpixels using SLIC [1]. Let $\mathbf{Y}^k = \{Y_1^k, Y_2^k, \cdots\}$ be the superpixel set of frame F^k. Given the pixel edge map E_c^k, the edge probability of each superpixel Y_n^k is computed as the average value of the pixels with ten largest edge probabilities within Y_n^k. This generates a superpixel edge map$\widehat{E}_c^k$. Similarly, we compute a superpixel optical flow magnitude map $\widehat{E}_o^k$ using E_o^k. Then a spatiotemporal edge probability map E^k is generated as:

$$\mathrm{E}^k = \widehat{E}_c^k \cdot \widehat{E}_o^k. \tag{1}$$

The intuition behind the design of (1) is that, if the motion patterns of foreground object distinct from background, the gradient of optical flow should have large magnitude around the object boundary. Additionally, the static edge maps give an instructor for the object boundaries according to the spatial information. When spatial edge and temporal discontinuity in motion are fused together through (1), the output spatiotemporal edges maps are able to imply the location of foreground object. This phenomenon could be easily observed from Fig. 1, the object regions either have high spatiotemporal edge values or are surrounded by these high-edge-probability regions. Based on this argument, we opt to use the geodesic distance to discriminate the visually salient regions from backgrounds and measure their likelihoods for foreground.

Intra-frame graph construction For frame F^k, we construct an undirected weighted graph $\mathcal{G}^k = \{\mathcal{V}^k, \mathcal{E}^k\}$ with superpixels $\mathbf{Y}^k$ as nodes $\mathcal{V}^k$ and the links between pairs of nodes as edges $\mathcal{E}^k$. The weight w_{mn}^k of the edge $e_{mn}^k \in \mathcal{E}^k$ between adjacent superpixels Y_m^k and Y_n^k is defined as:

$$e_{mn}^k = \|\mathrm{E}^k(Y_m^k) - \mathrm{E}^k(Y_n^k)\|, \tag{2}$$

where $\mathrm{E}^k(Y_m^k)$ and $\mathrm{E}^k(Y_n^k)$ correspond to the spatiotemporal boundary probability of superpixels Y_m^k and Y_n^k, separately. Based on the graph structure, we derive an $|\mathcal{V}^k| \times |\mathcal{V}^k|$ weight matrix W^k, where $|\mathcal{V}^k|$ is the number of nodes in $\mathcal{V}^k$. The (m, n)th element of W^k is: $W^k(m,n) = e_{mn}^k$. For each superpixel Y_n^k, the probability P_n^k for foreground is computed by the shortest geodesic distance to the image boundaries using

$$P_n^k = \min_{T \in \mathbf{T}^k} d_{geo}(Y_n^k, T, \mathcal{G}^k), \tag{3}$$

where $\mathbf{T}^k$ indicate the superpixels along the four boundaries of frame F^k. The geodesic distance $d_{geo}(v_1, v_2, \mathcal{G}^k)$ between any two superpixels $v_1, v_2 \in \mathcal{V}^k$ in graph $\mathcal{G}^k$ is defined as the accumulated edge weights along their shortest path on graph $\mathcal{G}^k$:

$$d_{geo}(v_1, v_2, \mathcal{G}^k) = \min_{C_{v_1,v_2}} \sum_{p=0,1} |W^k \cdot \dot{C}_{v_1,v_2}(p)|, \tag{4}$$

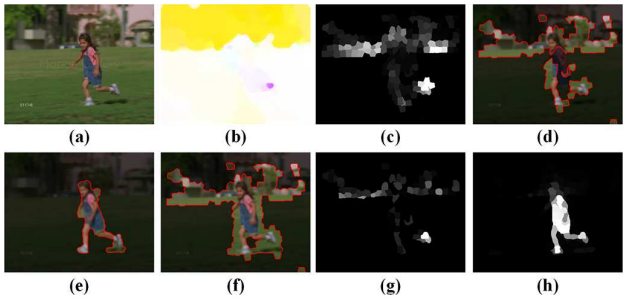

Figure 2. Illustration of inter-frame graph construction. (a) Frame F^k. (b) Optical flow flied V^k of (a). When the estimation for optical is not accurate (which is the usual case) object detection suffers P^k in (c). (d) Frame F^k is decomposed into background regions $\mathbf{B}^k$ and object-like regions $\mathbf{U}^k$ by self-adaptive threshold σ^k. The black regions indicate the background regions $\mathbf{B}^k$, while the bright regions indicate the object-like regions $\mathbf{U}^k$. (e) The decomposition of prior frame F^{k-1}. (f) The object-like regions $\mathbf{U}^{k-1}$ of frame F^{k-1} are projected onto frame F^k. (g) Spatiotemporal saliency result S^k for frame F^k with consideration of (d) and (e). (h) Spatiotemporal saliency result S^k for frame F^k with consideration of (e) and (f).

where $C_{v_1,v_2}(p)$ is a path connecting the nodes v_1, v_2 (for $p = 0$ and $p = 1$ respectively). If a superpixel is outside the desired object, its foreground probability is small because there possibly exists a pathway to image boundaries that does not pass the regions with high spatiotemporal edge value. Whereas, if a superpixel is inside the object, this superpixel is surrounded by the regions with large probabilities of edges, which increases the geodesic distance to image boundaries. We normalize all the foreground object probabilities P_n^k to [0, 1] for each frame, the object probability map for frame F^k is indicated by P^k. As our graph is very sparse, the shortest paths of all superpixels are efficiently computed by Johnson algorithm.

2.2. Spatiotemporal saliency

The obtained foreground probability map P^k can locate the foreground object but not very precisely. In particular, object probabilities of the background regions near the object boundaries are needless increased, due to the oversegmentation. Furthermore, erroneous results may come from the inaccuracy of optical flow estimation. Fortunately, foreground and background are visually different (by definition of saliency) and object is temporally continuous between adjacent frames. We present here a method which leverages this information to obtain spatiotemporal saliency results and is processed between pairs of adjacent frames.

Inter-frame graph construction For each pair of subsequent frame F^k and F^{k+1}, an undirected weighted graph $\mathcal{G}'^k = \{\mathcal{V}'^k, \mathcal{E}'^k\}$ is constructed. The nodes $\mathcal{V}'^k$ consist of all the superpixels $\mathbf{Y}^k$ of frame F^k and all the superpix-

els $\mathbf{Y}^{k+1}$ of frame F^{k+1}. There are two types of edges: intra-frame edges link all the spatially adjacent superpixels and inter-frame edges connect all the temporally adjacent superpixels. The superpixels are spatially connected if they are in the same frame and are adjacent, temporally adjacent superpixels refer to the superpixels which belong to different frames but have overlaps along the time axis. We assign the edge weight as the Euclidean distance between their average colors in the CIE-Lab color space.

For each frame, a self-adaptive threshold is used to decompose frame F^k into background regions $\mathbf{B}^k$ and object-like regions $\mathbf{U}^k$ through the object probability map P^k. This threshold σ^k for frame F^k is computed by $\sigma^k = \mu(P^k)$, where $\mu(\cdot)$ computes the mean probability of all pixels within frame F^k by probability map P^k. Additionally, the background information of previous frame offers valuable prior, which could eliminate the artifacts due to the inaccurate optical flow estimation. Therefore, we define the background regions $\mathbf{B}^k$ of k-th frame as:

$$\begin{aligned}\mathbf{B}^k = &\ \{Y_n^k | P_n^k \leq \sigma^k\} \\ &\cup \{Y_n^k | Y_n^k \text{ is temporally connected to } \mathbf{B}^{k-1}\}, \\ \mathbf{U}^k = &\ \mathbf{Y}^k - \mathbf{B}^k,\end{aligned} \tag{5}$$

Based on the graph ${\mathcal{G}'}^k$, we obtain a saliency value S_n^k (P_n^{k+1}) of superpixels Y_n^k (Y_n^{k+1}) of frame F^k (F^{k+1}) as follows:

$$S_n^k = \min_{B \in \mathbf{B}^k \cup \mathbf{B}^{k+1}} d_{geo}(Y_n^k, B, {\mathcal{G}'}^k). \tag{6}$$

The main rationale behind the relation in (6) is that a saliency value of a superpixel is measured by its shortest path to background regions in color space, both considering spatial and temporal background information. Fig. 2 gives illustration of this process. After obtaining spatiotemporal saliency map S^k and S^{k+1} for frame F^k and F^{k+1} , we keep executing this process for next two adjacent frame F^{k+1} and F^{k+2} until the end of the video sequence.

2.3. Spatiotemporal object segmentation

We formulate video object segmentation as a pixel labeling problem with two labels (foreground and background). Each pixel $x_i^k \in \mathbf{X}^k$ can take a label $l_i^k \in \{0, 1\}$, where 0 corresponds to background and 1 corresponds to foreground. A labelling $\mathbf{L} = \{l_i^k\}_{k,i}$ of pixels from all frames represents a segmentation of the video. Similarly to other segmentation works [15, 27], we define an energy function for labeling $\mathbf{L}$ of all the pixels:

$$\begin{aligned}\mathcal{F}(\mathbf{L}) = &\sum_{k,i} \mathcal{U}_i^k(l_i^k) + \lambda_1 \sum_{k,i} \mathcal{A}_i^k(l_i^k) + \lambda_2 \sum_{k,i} \mathcal{L}_i^k(l_i^k) \\ &+ \lambda_3 \sum_{(i,j)\in \mathbf{N}_s} \mathcal{V}_{ij}^k(l_i^k, l_j^k) + \lambda_4 \sum_{(i,j)\in \mathbf{N}_t} \mathcal{W}_{ij}^k(l_i^k, l_j^{k+1}),\end{aligned} \tag{7}$$

where spatial pixel neighborhood $\mathbf{N}_s$ consists of eight spatially neighboring pixels within one frame, temporal pixel neighborhood $\mathbf{N}_t$ consists of the forward-backward nine neighbors in adjacent frames, and i, j index the pixels.

This energy function consists of three unary terms, $\mathcal{U}^k$, $\mathcal{A}^k$ and $\mathcal{L}^k$, and two pairwise terms $\mathcal{V}^k$ and $\mathcal{W}^k$, which depend on the labels of spatially and temporally neighboring pixels. The scalar parameters λ weight the various terms. In our experiments, we set $\lambda_1 = \lambda_2 = 0.5, \lambda_3 = \lambda_4 = 4$. The purpose of $\mathcal{U}^k$ is to evaluate how likely a pixel is foreground or background according to saptio-temporal saliency maps computed by prior step. The unary appearance term $\mathcal{A}^k$ encourages labeling pixels which have similar colors as pixels with high saliency for foreground. The third unary term $\mathcal{L}^k$ is defined for labeling pixels with location priors estimated from dynamic location models. The pairwise terms $\mathcal{V}^k$ and $\mathcal{W}^k$ encourage spatial and temporal smoothness, respectively. All the terms are described in detail next.

Saliency term $\mathcal{U}^k$. The unary saliency term $\mathcal{U}^k$ is based on our saliency detection results, which penalizes labelings which assign pixel with low saliency value to the foreground. The term $\mathcal{U}^k$ has the following form:

$$\mathcal{U}^k(l_i^k) = \begin{cases} -\log(1 - S^k(x_i^k)) & \text{if } l_i^k = 0; \\ -\log(S^k(x_i^k)) & \text{if } l_i^k = 1. \end{cases} \tag{8}$$

Appearance term $\mathcal{A}^k$. To model the foreground and background appearance, two weighted color histograms are computed in RGB color space, which should be denoted by H_f and H_b. Each color channel is uniformly quantized into 10 bins, and there is a total of 10^3 bins. Each pixel is stacked into histograms according to its color values and weighted by its saliency value, where the weight for pixel x is $S^k(x)$ and $1 - S^k(x)$ for H_f and H_b, respectively. Then we establish global appearance model for foreground and background by normalizing H_f and H_b.

More specially, pixels belonging to two kinds superpixels are sampled for forming H_f and H_b: one that the superpixels with saliency value larger than the adaptive threshold defined as the mean value of spatiotemporal saliency map, and one that the superpixels spatially connected to the former superpixels. We denote these pixels as $\mathbf{X}_s$. This strategy makes full use of the information of spatiotemporal saliency results and is able to eliminate ill effects of some background regions with similar color to the foreground, thus offering more accurate fore-/background estimation. Let $c(x_i^k)$ denotes the histogram bin index of RGB color value at pixel x_i^k, the unary appearance term $\mathcal{A}^k$ is defined

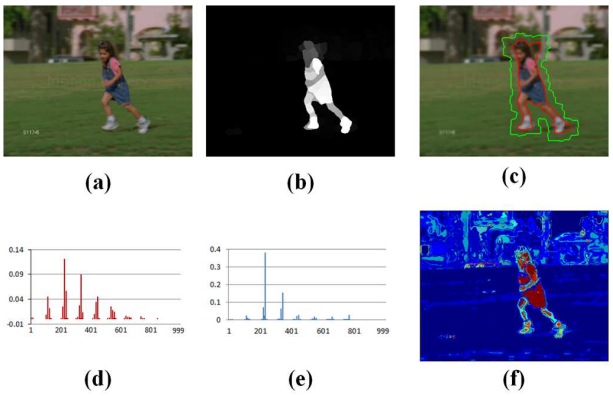

Figure 3. The illustration of establishing appearance model. (a) Input frame F^k. (b) Spatiotemporal saliency map S^k. (c) Pixel set $\mathbf{X}_s$ for the frame in (a), which consist of the pixels within the green boundaries. The regions within the red boundaries are the superpixels with the saliency value larger than the adaptive threshold. (d)-(e) The global appearance model with color histogram H_f (d) and H_b (e) for foreground and background respectively, which are sampled from all the pixels belonging to $\mathbf{X}_s$ for each frame. (f) The probability map for foreground computed via global appearance model.

as:

$$\mathcal{A}^k(l_i^k) = \begin{cases} -\log(\frac{H_b(c(x_i^k))}{H_f(c(x_i^k)) + H_b(c(x_i^k))}) & \text{if } l_i^k = 0; \\ -\log(\frac{H_f(c(x_i^k))}{H_f(c(x_i^k)) + H_b(c(x_i^k))}) & \text{if } l_i^k = 1. \end{cases} \tag{9}$$

Location term $\mathcal{L}^k$. Even above efforts for making the appearance model as accurate as possible pay off, the estimation can still be distorted when the scene is complex or the background regions share similar appearance with foreground. To this, the object motion continuity among few subsequent frames, provides a valuable prior to locate the areas likely to contain the object. Thus, we design a method to estimate location of foreground object with with respect to motion information from a small number of neighboring frames. For k-th frame, we accumulate its forward-backward t frames' optical flow gradient magnitude that yields trajectory of the object within few subsequent frames:

$$E_t^k = \sum_{i=k-t}^{k+t} E_o^i = \sum_{i=k-t}^{k+t} \|\nabla V^i\|. \tag{10}$$

Having a larger t for a certain frame, long-range motion information will be taken into account ignoring some unreliable optical flow estimation from small number of frames. However, this possibly makes E_t^k lose discriminative ability for object since too much motion information is unnecessary. When t is as small as 0, only considering current frame's motion information possibly precisely prime the object location but sometimes will fail because of inaccurate optical flow estimation. In our experiments, we set $t = 5$. Then we use the within-frame graph construction method described in section 2.1 to compute a dynamic location model for each frame. Finally, we can get location prior L_i^k for pixel x_i^k, and the unary location term $\mathcal{L}^k$ is defined as:

$$\mathcal{L}^k(l_i^k) = \begin{cases} -\log(1 - L^k(x_i^k)) & \text{if } l_i^k = 0; \\ -\log(L^k(x_i^k)) & \text{if } l_i^k = 1. \end{cases} \tag{11}$$

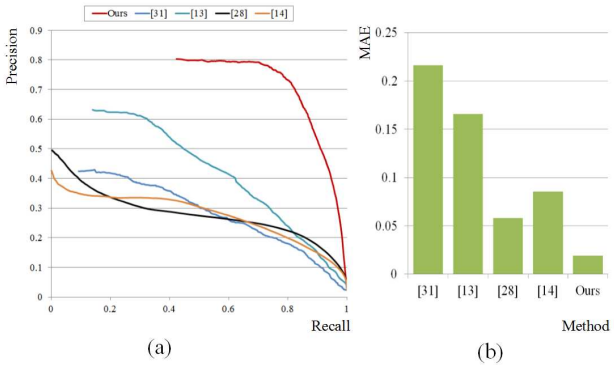

Figure 4. Statistical comparison with 5 alternative saliency detection methods using SegTrack dataset [29] with pixel-level ground truth: (a) average precision recall curve by segmenting saliency maps using fixed thresholds, (b) average MAE. Notice that our algorithm significantly outperforms other methods in terms of the precision-recall. Additionally, our method achieved 75% improvement over the best previous method in terms of MAE.

Pairwise terms $\mathcal{V}^k$, $\mathcal{W}^k$. $\mathcal{V}^k$, $\mathcal{W}^k$ compose the consistency term, constraining the segmentation labels to be both spatially and temporally consistent. These two terms follow the conventional form defined in [27], which favors assigning the same label to neighboring pixels that have similar color.

Having defined the complete energy function $\mathcal{F}(\mathbf{L})$, we can use graph-cuts to compute the optimal binary labeling, and thus get the final segmentation results.

3. Experimental results

Our approach automatically detects and segments the foreground object in the video sequences. In this section, we first test our method on video saliency detection. Even though it is not the final goal of our proposed algorithm, we still evaluate the effectiveness of our approach by comparing our spatiotemporal saliency results against the state-of-art saliency methods [31, 13, 28, 14] on the SegTrack dataset [29]. Then we compare our segmentation results with 9 alternate methods on the SegTrack [29], SegTrack v2 [18] and Youtube datasets.

Figure 5. Comparison of previous methods to our spatiotemporal saliency results using SegTrack dataset [29] with ground truth (GT).

method	Ours	[32]	[23]	[15]	[6]	[19]	[4]	[22]	[29]	[9]
birdfall	209	155	189	288	217	468	468	606	252	454
cheetah	796	633	806	905	890	1175	1968	11210	1142	1217
girl	1040	1488	1698	1785	3859	5683	7595	26409	1304	1755
monkeydog	562	365	472	521	284	1434	1434	12662	563	683
parachute	207	220	221	201	855	1595	1113	40251	235	502
Avg.	427	452	542	592	868	1727	1911	19079	594	791
supervised	N	N	N	N	N	N	N	N	Y	Y

Table 1. The average per-frame pixel error rate using SegTrack dataset [29] compared to the ground-truth .

In the proposed algorithm, we utilize the spatiotemporal edge information to compute the prior saliency maps for videos. As this is an important step of our method, we evaluate the results through other saliency methods. Using the codes obtained from the corresponding authors, we compare our spatiotemporal saliency results with five alternate methods [31, 13, 28, 14]. The first method aims at image saliency detection while the later three ones are designed for video saliency detection. To evaluate the performance of our method, we test our results based on two widely used criteria, including PR (precision-recall) curve and MAE (mean absolute errors). We first evaluate our method using precision recall analysis. $Precision$ is defined as the percentage of salient pixels correctly assigned, while $recall$ measures the percentage of salient pixel detected. To plot the precision-recall cures, we generate binary saliency maps from each method using a fixed threshold. The PR curve is drawn by 256 precision-recall pairs, which are obtained by varying the threshold from 0 to 255.

For a more balanced comparison, we follow Perazzi *et al.* [24] to evaluate the $mean\ absolute\ error$ (MAE) between a continuous saliency map $\mathbb{S}$ and the binary ground truth $\mathbb{G}$ for all image/frame pixels. MAE is defined as: MAE $= |\mathbb{S} - \mathbb{G}|/N$, where N is the number of image/frame pixels. The MAE estimates the approximation degree between the saliency map and the ground truth, which is normalized to [0, 1]. MAE provides a better estimate of dissimilarity between the saliency map and ground truth.

The resulting precision recall curve is illustrated in Fig. 4(a), which provides a reliable comparison of how well var-

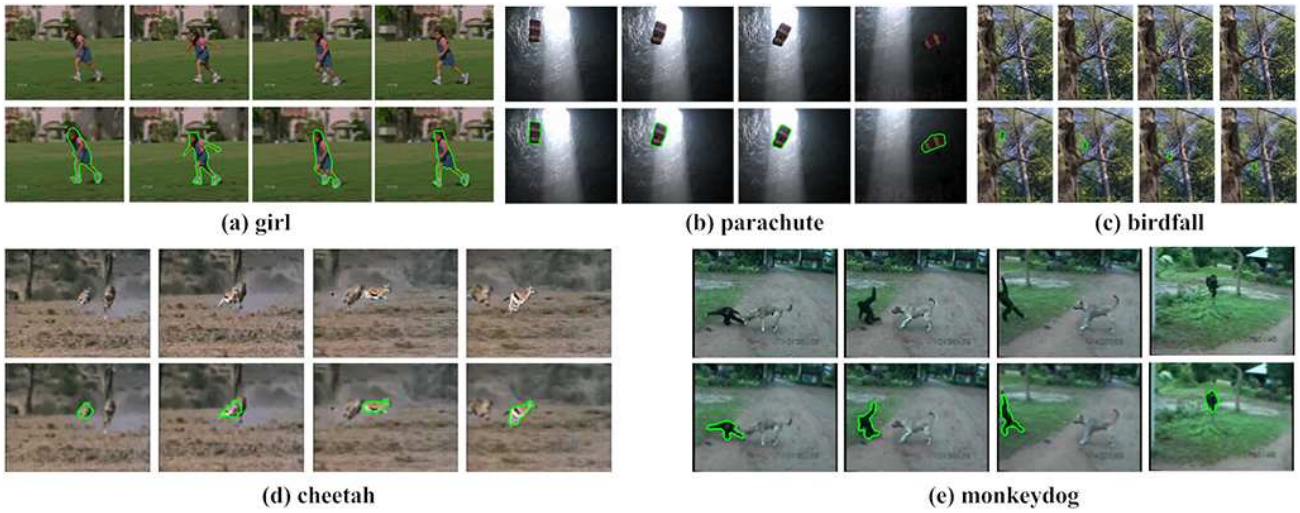

Figure 6. Our segmentation results using SegTrack dataset[29]. The regions within the green boundaries are the segmented foreground objects.

ious saliency maps highlight salient regions in images. The results show that the proposed algorithm significantly outperforms other methods. When the threshold is close to 255, the recall values of [31, 13, 28, 14] are very small, even the recall values of [28] and [14] decrease to 0, since their saliency maps do not respond to the objects of attention. The minimum recall value of the proposed method does not drop to zero because the corresponding saliency maps are able to effectively detect the salient region with strong response. Moreover, our saliency method achieves the best performance up to a precision rate above 0.8, which indicates our saliency maps are more precise and responsive to the salient regions. The MAE results are presented in Fig. 4(c). Our saliency maps successfully reduce the MAE by 75% compared to the best result [29] of other methods.

Fig. 5 gives a visual comparison of different methods, where brighter pixels indicate higher saliency probabilities. The performance of image saliency method [31] is not well, some saliency maps even cannot correctly detect the foreground object. The lack of motion information limits their ability to precisely localize object, especially when foreground and background have similar color. In most cases, saliency methods [13, 28, 14] for video are able to accurately locate the salient objects, which perform better than the method [31] for image saliency detection. Since those spatiotemporal methods utilize motion information. However, some saliency maps using [13, 28, 14] are generated in low resolution and tend to assign relatively low probabilities to pixels inside the objects. That is because optical flow estimation sometimes is not correct. Based on prior analysis, we can draw two important conclusions: (1) motion information gives effective guidance for detecting foreground object; (2) making methods excessively dependent on motion information is not an excellent choice. Comprehensive utilization of various features in spatial and temporal space (etc. color, edges, motion) should produce more satisfied results. Overall, our model is able to better estimated saliency maps at pixel level within and on the contour of the objects in cluttered backgrounds.

Our framework produces both spatially and temporally coherent object segmentation results for videos in a fully unsupervised way, and we compare nine methods that are the most closely related works published in recent years. The average per-frame pixel error rate [29] is introduced for evaluation, which is the number of pixels misclassified according to the ground truth segmentation. The average per-frame pixel error rate compared with these methods [32, 23, 15, 6, 19, 4, 22, 29, 9] for each video from SegTrack dataset [29] are summarized in Table 1. The methods in [32, 23, 15, 6, 19, 4, 22] and our method are unsupervised. They automatically detect object in video as well as segment the object out. The methods in [29] and [9] are supervised, which require an initial annotation for the first frame. As the results shown, our method has the lowest average per frame segmentation error over the test videos.

Fig. 6 shows qualitative results for the videos of SegTrack dataset [29]. It can be observed that our method has the ability to segment the objects with large shape deformation (*girl*), foreground/background color overlap (*parachute*) and camera motion (*monkeydog*), and also produces accurate segmentation even when the objects are very small (*birdfall*), or foreground with fast motion patterns (*cheetah*).

We further carried out experiments on SegTrack v2 dataset [18] and 12 groups of videos randomly selected from Youtube Objects and compared our method with [32, 23, 15, 6] as well. The average per-frame pixel error rate are illustrated in Table 2. As seen, our method sig-

dataset	Ours	[32]	[23]	[15]	[6]
SegTrack v2	4766	25289	5859	23161	16074
Youtube	2208	11148	3461	20115	16858

Table 2. The average per-frame pixel error rate using SegTrack v2 [18] and Youtube dataset compared to the ground-truth .

nificantly outperforms all others on SegTrack v2 [18] and Youtube dataset too.

4. Conclusions

We presented an unsupervised method that incorporates geodesic distance into saliency empowered video object segmentation. The proposed spatiotemporal edge map is shown to be able to indicate the location of foreground and background. Our approach integrated spatiotemporal edge map and geodesic distance to obtain accurate spatiotemporal saliency results as a prior to object segmentation. We produced spatiotemporal saliency maps via the computation of geodesic distance to the estimated background on the inter-frame graph for each pair of adjacent frames. Finally, we computed the segmentation result by combining saliency, global appearance model and location model into the graph-cut energy minimization. Numerous results showed that our approach yields clearly higher performance than the state-of-the-art methods.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*, 34(11), 2012.

[2] C. Antonio, S. Toby, and B. Andrew. Geos: geodesic image segmentation. In *ECCV*, 2008.

[3] X. Bai and G. Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *ICCV*, 2007.

[4] O. Barnich and M. Van Droogenbroeck. Vibe: a universal background subtraction algorithm for video sequences. *IEEE TIP*, 20(6), 2011.

[5] W. Brendel and S. Todorovic. Video object segmentation by tracking regions. In *CVPR*, 2009.

[6] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[7] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE T-PAMI*, 33(3), 2011.

[8] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010.

[9] P. Chockalingam, N. Pradeep, and S. Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. In *CVPR*, 2009.

[10] A. Criminisi, T. Sharp, C. Rother, and P. Perez. Geodesic image and video editing. *ACM TOG*, 29(5), 2010.

[11] I. Endres and D. Hoiem. Category independent object proposals. In *ECCV*, 2010.

[12] K. Fragkiadaki, G. Zhang, and J. Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012.

[13] H. Fu, X. Cao, and Z. Tu. Cluster-based co-saliency detection. *IEEE TIP*, 22(10), 2013.

[14] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, 2008.

[15] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011.

[16] M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Efficient closed-form solution to generalized boundary detection. In *ECCV*, 2012.

[17] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011.

[18] F. Li, T. Kim, A. Humayun, D. Tsai, and J. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013.

[19] T. Ma and L. J. Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, 2012.

[20] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE TPAMI*, 32(1), 2010.

[21] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *ECCV*. 2012.

[22] P. Ochs and T. Brox. Higher order motion models and spectral clustering. In *CVPR*, 2012.

[23] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[24] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.

[25] B. Price, B. Morse, and S. Cohen. Geodesic graph cut for interactive image segmentation. In *CVPR*, 2010.

[26] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting salient objects from images and videos. In *ECCV*, 2010.

[27] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM TOG*, 23(3), 2004.

[28] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12), 2009.

[29] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg. Motion coherent tracking using multi-label mrf optimization. In *IJCV*, 2012.

[30] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *ECCV*, 2012.

[31] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[32] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013.