Video Co-saliency Guided Co-segmentation

Wenguan Wang, Jianbing Shen, *Senior Member, IEEE*, Hanqiu Sun, and Ling Shao, *Senior Member, IEEE*

***Abstract*—We introduce the term *video co-saliency* to denote the task of extracting the common noticeable, or salient, regions from multiple relevant videos. The proposed video co-saliency approach accounts for both inter-video foregrond correspondences and intra-video saliency stimuli to emphasize the salient foreground regions of video frames and, at the same time, disregard irrelevant visual information of the background. Compared to image co-saliency, it is more reliable due to the utilization of temporal information of video sequence. Benefiting from the discriminability of video co-saliency, we present a unified framework for segmenting out common salient regions of relevant videos, guided by video co-saliency prior. Unlike naive video co-segmentation approaches employing simple color differences and local motion features, the presented video co-saliency provides a more powerful indicator for the common salient regions, thus conducting video co-segmentation efficiently. Extensive experiments show that the proposed method successfully infers video co-saliency and extracts the common salient regions, outperforming the state-of-the-art methods.**

***Index Terms*—Video co-saliency, video co-segmentation, video salient region.**

I. Introduction

With the fast growth of video data, leveraging such resources for learning and making it accessible and searchable in an easy way is an interesting research topic in computer vision. Video collections, which typically share common objects, exhibit some important properties, such as the object class structure and object relations. These properties offer more useful and rich information about the object, which leads to a joint consideration rather than processing each video independently. The frequently occurred patterns in video collections usually capture more human attention and make the content more understandable. Noticing this important fact, we explore the task of identifying and segmenting out the common salient regions from a group of related videos. The latter is known as *video co-segmentation* while the former we term *video co-saliency*. Video co-saliency aims to extract the common saliency from multiple videos with weak annotation as sharing similar salient objects. The outcome is a group of saliency maps where the intensity of each pixel represents the probability of that pixel belonging to the common salient regions. We offer a detailed definition of video co-saliency, which simultaneously exhibits two basic properties: 1) video co-saliency should follow the law of visual saliency within an individual video sequence, which is discriminative from the background; 2) video co-saliency should be similar across multiple videos, which can be identified as the common patterns via leveraging the global repetitiveness distribution. The latter definition is potentially useful in various applications, including video object co-localization and co-recognition, and also contributes to our video co-segmentation task.

Visual saliency is a key attentional mechanism that facilitates learning by focusing limited perceptual and cognitive resources on the most pertinent subset of the available visual data. Saliency detection [1], [2], [3], [4], [5], [6] for static scenes has gained much attention due to its importance in many computer visual tasks. Compared with image saliency detection, relatively few works [7], [8], [9], [10], [11] address the problem for video sequences. In dynamic scenes, moving objects are more likely to grab human attention. Only recently, image co-saliency [12], [13], [14], [15], [16], [17] has been introduced to discover the common saliency on multiple images. Nonetheless, co-saliency for multiple videos is still an under-investigated area. Compared with static and dynamic saliency detection, video co-saliency takes advantage of the appearance and structure information of the foreground across multiple videos to determine the prevalent salient regions.

Video co-segmentation aims at jointly segmenting common objects from two or more videos [18], [19], [20], [21], [22]. Compared with object segmentation from a single video [23], [24], [25], the benefit is that object correspondences across multiple videos are leveraged for segmentation. Video co-saliency serves as a means of compensating for the lack of supervisory data, naturally contributing to video co-segmentation. Building on the observation that common objects should also be salient across multiple relevant videos, video co-saliency offers a strong indicator of common salient object. Our source code will be available at [1].

This paper can be viewed as a very early work for video co-saliency and exploring the use of video co-saliency for conducting video co-segmentation. The key contributions of this paper are:

- A new video co-saliency approach is presented for deriving the common saliency from multiple relevant videos that are weakly annotated for sharing similar salient foregrounds.
- A new video co-segmentation method incorporating video co-saliency prior, which offers a powerful indicator for

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 RGC research grant (No. 416212), the UGC direct grant for research (No. 4055060), and the Fok Ying-Tong Education Foundation for Young Teachers. Specialized Fund for Joint Building Program of Beijing Municipal Education Commission. (Corresponding author: Jianbing Shen)

W. Wang and J. Shen are with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, P. R. China. (Email: shenjianbing@bit.edu.cn)

H. Sun is with Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. (Email: hanqiu@cse.cuhk.edu.hk)

L. Shao is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K. (Email: ling.shao@ieee.org)

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

the common salient regions, is introduced in a unified framework.
- Various discriminative saliency cues for video co-saliency and a saliency cues integration method are proposed and we show the proposed method achieves promising results on well-known datasets.

II. Related work

A. Saliency Detection

Image saliency detection, attempting to identify the most informative and important regions in static scenes, can be categorized as either top-down or bottom-up approaches. Top-down methods [26], [27] are task-driven and usually adopt supervised learning with a specific class. Bottom-up methods are often rely on heuristic assumptions about the properties of salient regions due to the absence of high level knowledge. The most widely used assumption, called *contrast prior*, is that a pixel (or a patch) is salient if its appearance is high contrast within a certain context. Almost all saliency methods [2], [28] build their saliency models on such an assumption. Besides contrast prior, several recent works [1], [3], [29] formulate *boundary prior*, *i.e.,* image boundary regions are more likely to be the background, for further enhancing saliency computation.

Some efforts were paid for detecting saliency from video sequences. Some methods [28], [30] combined image saliency model with additional motion channel. Other methods used local regression kernels [8], statistical model [31], or various low-level saliency features [9]. Wang *et al.* [32], [33] proposed a geodesic distance based video saliency method and employed such saliency prior to guide video object segmentation.

There are relatively few works [12], [13], [14], [15] that explore co-saliency for still images. Those approaches utilize additional companion images as cues to discover the common saliency. Li *et al.* [12] designed a single-image saliency cue and a multi-image saliency cue for inferring the common saliency from a pair of images. In [13], the regions which are salient in a single view and frequently repeat in most images are emphasized. In [14], contrast and spatial cues are combined with a corresponding cue for discovering the common salient objects on the multiple images using clustering distribution. Cao *et al.* [15] exploited the relationship of multiple saliency cues and adopted a self-adaptive weight to generate the final saliency/co-saliency via the rank constraint. We refer reader to a recent survey [34] for more details of image co-saliency.

B. Unsupervised Video Segmentation

Separating objects from the background in a video sequence is a classical problem in computer vision. A large number of unsupervised methods have been proposed for this problem in the past decades. Some methods [35], [36], [37] tracked feature points or local regions over frames and clustered the tracks for extracting the moving objects. Many methods [23], [24], [31], [38] favored utilizing both motion and appearance cues. Some efforts [39], [40], [41] were paid to solve this problem based on the notion of what a generic object looks like. They usually generate a large number of object proposals in every frame, and select the most relevant object proposal for generating final segments. In the selection process, both motion and appearance cues are combined in measuring the *objectness* of proposals where various assessment strategies are introduced. More recently, some methods [32], [42] were proposed to utilize spatiotemporal saliency for guiding segmentation.

C. Video Co-Segmentation

Video co-segmentation is for simultaneously segmenting a common category of objects from a group of related videos. The idea of video co-segmentation was first introduced by Rubio *et al.* [43]. Since then, some works [19], [20], [21] extracted multiple foreground objects by a non-parametric Bayesian model or object proposals. Guo *et al.* [18] applied video co-segmentation for common action extraction using dense trajectories. Wang *et al.* [22] proposed a spatiotemporal SIFT flow for establishing correspondences among videos and thus inferring the common objects. Some efforts [44], [45], [46], [47] were paid for discovering the common foreground from a group of similar videos/images. Some methods were proposed for unsupervised object discovery [48], [49], which are proceeded on unconstraint video groups.

III. Our Approach

Fig. 1 shows the flowchart of our method. The proposed method can be decomposed into two main stages: video co-saliency estimation (Sec. III-A) and video co-segmentation (Sec. III-B). In the initial video co-saliency stage, we utilize visual properties across multiple videos to infer the common salient regions, accounting for inter-video saliency coherence and intra-video saliency stimuli. Three saliency cues: inter-video saliency, region-contrast saliency, and location saliency, are integrated to generate the video co-saliency. In the video co-segmentation stage, an energy optimization formulation that uses a Markov Random Field model to incorporate the co-saliency cues generated by the prior step is employed. This co-segmentation energy function also utilizes single-video and multi-video appearance models for the foreground and background. Next, we present the video co-saliency and co-segmentation stages in detail.

A. Video Co-saliency

The task of video co-saliency is to find the common salient regions from related videos. Suppose the set of videos are $\mathcal{V} = \{V^1, ..., V^N\}$, where each video V^n consists of frames $\mathcal{I}^n = \{I_1^n, ..., I_t^n, ...\}$. In each frame I_t^n, a set of superpixels X_t^n is obtained by [50], due to its superiority in terms of adhering to boundaries, as well as computational efficiency. Superpixels in video V^n are then $\mathcal{X}^n = \{X_1^n \cup ... \cup X_t^n \cup ...\}$.

For the first frame I_1^n $(t = 1)$ within V^n, we produce an initial saliency map $S_{n,1}^I$ by averaging the outputs of various saliency methods [1], [2], [3] as an initial estimation for foreground and background. These three approaches are based on different saliency priors. [1] utilizes *boundary* prior and [2] considers *contrast* prior mainly. We use [3] as it both considers boundary prior and contrast prior.

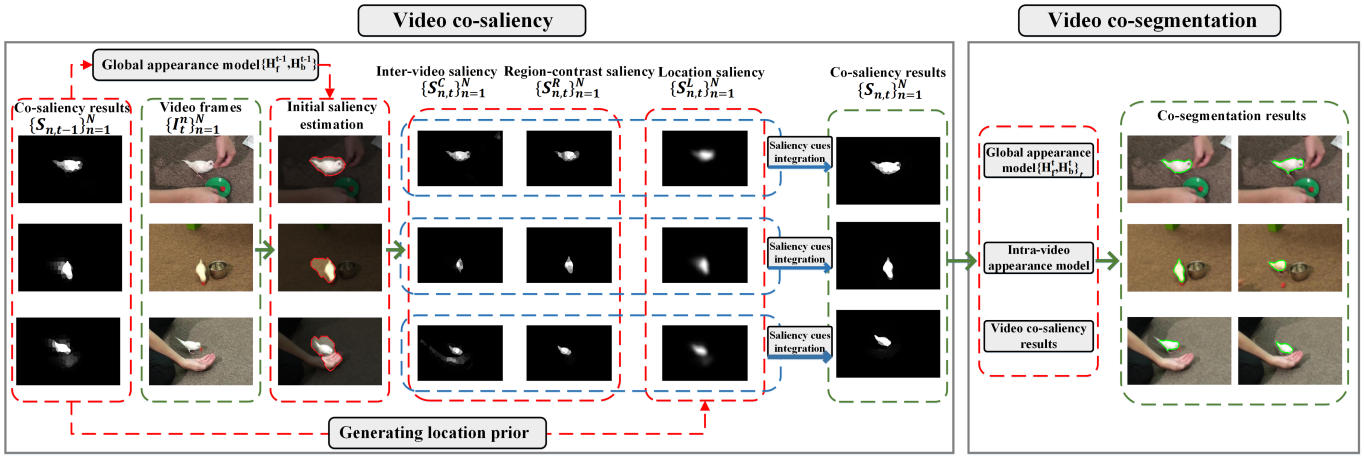


Fig. 1. Flowchart of our method. We use co-saliency results $\{S_{n,t-1}\}_{n=1}^{N}$ of $\{t–1\}$-th frames to establish initial foreground/background estimation for t-th frames $\{I_t^n\}_{n=1}^{N}$ for all the videos. $\{\mathcal{B}_t^n\}_{n=1}^{N}$ (dark) indicates the background area while $\{\mathcal{F}_t^n\}_{n=1}^{N}$ (bright) corresponds to the salient region. Based on this, we further introduce an inter-video saliency cue S^C and a self-adapting region-contrast saliency S^R. Additionally, a refined location prior S^L is derived from $\{S_{n,t-1}\}_{n=1}^{N}$ and all saliency cues are integrated for the final co-saliency $\{S_{n,t}\}_{n=1}^{N}$ for t-th frames in all videos. Finally, a co-segmentation energy function considering co-saliency cue, global appearance model and intra-video appearance model is employed to segment out common salient regions.

For the frame I_t^n $(t > 1)$, the initial saliency map $S_{n,t}^I$ is produced by an inter-video appearance model estimated from previous co-saliency results. Since the initial foreground/background estimation relies on the previous frame's co-saliency results, we first describe the process for generating co-saliency results from these initial saliency maps. Then we introduce how to generate these initial saliency maps from the prior frame's co-saliency results.

With initial saliency estimation for frame I_t^n, we can approximately separate I_t^n into two parts: salient region and background area. A simple thresholding method can achieve this goal. A superpixel is labeled as salient if its initial saliency value exceeds a certain threshold. A binary map M_t^n is obtained on the initial saliency $S_{n,t}^I$:

$$M_t^n(x) = \begin{cases} 1 & \text{if } S_{n,t}^I(x) \le \tau; \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $x \in X_t^n$, and $S_{n,t}^I(x)$ is the saliency value of the superpixel x by the initial saliency map $S_{n,t}^I$. The background area $\mathcal{B}_t^n$ (dark pixels in Fig. 1) is defined as a set of superpixels with small initial saliency values:

$$\mathcal{B}_t^n = I_t^n \cdot M_t^n, \tag{2}$$

We set the threshold value as $\tau = 0.2$ to exclude the salient region from the background as much as possible. Then, the region set $\mathcal{X}_t^n$ of frame I_t^n is decomposed into two parts: salient area $\mathcal{F}_t^n$ and background area $\mathcal{B}_t^n$. To identify the common foreground regions from the related videos $\mathcal{V}$, we need to highlight common foreground area and exclude the background part from the salient area $\{\mathcal{F}_t^n\}_{n=1}^{N}$. Aiming for this, we introduce various saliency cues that consider foreground coherence among different videos and object characteristics within single video.

1) Inter-video saliency cue: It is a valid observation that the appearance of foreground regions is often distinct from that of the background. Additionally, common salient regions will share similar appearances across different videos. Based on these observations, we use an inter-video saliency cue to measure the saliency using color statistics across multiple videos. Suppose H_t^f and H_t^b denote the normalized color histograms sampled from salient region $\{\mathcal{F}_t^n\}_{n=1}^{N}$ and background area $\{\mathcal{B}_t^n\}_{n=1}^{N}$. For the t-th frame I_t^n within each video V^n, the inter-video saliency measure is given by:

$$S_{n,t}^C(x) = \frac{H_t^f(x)}{H_t^f(x) + H_t^b(x)}, \tag{3}$$

where $x \in \mathcal{X}_t^n$ and $H_t^f(x)$ $(H_t^b(x))$ denotes the histogram value according to RGB color value of superpixel x. It is intuitive that frequently occurred regions among different video sequences easily attract human attention. This measure considers the appearance consistency of the salient regions across multiple videos, since the superpixels from all the salient regions and background areas are stacked into the histograms. A superpixel with more frequent colors across different salient regions will be assigned a higher saliency value. It is easy to see regions with the same color value across different videos share same saliency value under this definition, since the measure is oblivious to intra-video object indicators such as local contrast and spatial locations. For this reason, two intra-video saliency cues are further introduced into our model.

2) Self-adapting region-contrast saliency: We define a self-adapting region-contrast saliency, denoted by S^R, as the weighted sum of the superpixel's contrasts to all other background area in the frame. With the background area $\mathcal{B}_t^n$ for frame I_t^n, S^R is designed as:

$$S_{n,t}^R(x) = \sum_k \sum_{x_b \in \mathcal{B}_t^n} e^{-\frac{D_s(x,x_b)}{\ell(x)}} \| f_k(x) - f_k(x_b) \|_1, \tag{4}$$

where $\{f_k\}_k$ indicate features of superpixel and $k = \{c, m, o\}$ correspond to color, magnitude and orientation of optical flow. $D_s(\cdot,\cdot)$ is the Euclidean distance between the centroids of respective superpixels. In our work, the optical flow is computed

by LDOF [51]. Here, we use a self-adapting influence radius $\ell(x)$ to measure the strength of spatial weighting for superpixel x:

$$\ell(x) = \min_{x_b \in \mathcal{B}_t^n} D_s(x, x_b). \tag{5}$$

This influence radius $\ell(x)$ is designed as the shortest spatial distance from superpixel x to intra-frame background $\mathcal{B}_t^n$. If a superpixel is inside salient object, far away from the background, a large $\ell(x)$ would make far regions contribute more to the saliency of the current region. Conversely, if a superpixel spatially circumvents the background, the value of $\ell(x)$ is decreased and only the contrast to the background in a small area would be considered.

3) Location saliency cue: In addition to the contrast between a superpixel and its surrounding background, we also compute a location prior term considering spatial relationships. It is a fact that video production often frames the regions of interest near the center of the screen. Many methods incorporate this location bias to their saliency models by assigning a constant 2D Gaussian positioned at the center of a frame. Inspired by this, we further introduce a new location prior into our saliency estimation. Since a prior frame's saliency has been produced, we can get a more accurate location model instead of simply assuming the pixels with the same distance to the center of a frame share a saliency value. With $\{t-1\}$-th frame's saliency estimation $S_{n,t-1}$ of video V^n, we compute a location bias $S_{n,t}^L$ for frame I_t^n by:

$$S_{n,t}^L = \hbar(S_{n,t-1}), \tag{6}$$

where $\hbar(S_{n,t-1})$ describes average-filtering the saliency map $S_{n,t-1}$. In practice, we observed that a linear filter is preferable to a Gaussian filter, which falls off too sharply and the size of the smoothing window was set as 9×9.

4) Saliency cues integration: Given a group of relevant videos, the proposed video co-saliency method efficiently produces three saliency estimates, i.e. inter-video saliency S^C, region-contrast saliency S^R and location prior S^L. They are derived by using different information of video sequences, hence they are complementary and can be fused to further improve the performance.

To combine multiple saliency stimuli, instead of simply using a weighted linear combination of the saliency as previous works, we propose an energy function for intuitively merging different saliency cues and generating the final co-saliency. We start by normalizing S^C, S^R and S^L to the range of [0, 1]. We model the video co-saliency detection problem as the optimization of the co-saliency values of all superpixels. The objective energy function is designed to assign the final co-saliency value according to their initial co-saliency assignments and considers spatiotemporal consistency. The optimal saliency map is then obtained by minimizing the energy function:

$$\begin{aligned} E(X_t^n) = \; & \alpha_1 \sum_i E^C(x_i) + \alpha_2 \sum_i E^R(x_i) \\ & + \alpha_3 \sum_i E^L(x_i) + \sum_{i,j \in \aleph_s} E^S(x_i, x_j). \end{aligned} \tag{7}$$

The data terms $E^C(\cdot)$, $E^R(\cdot)$ and $E^L(\cdot)$ are to measure the cost between final co-saliency S with various saliency cues: S^C, S^R and S^L respectively, and hence are defined as:

$$\begin{aligned} E^C(x_i) &= (S_{n,t}(x_i) - S_{n,t}^C(x_i))^2 \\ E^R(x_i) &= (S_{n,t}(x_i) - S_{n,t}^R(x_i))^2 \\ E^L(x_i) &= (S_{n,t}(x_i) - S_{n,t}^L(x_i))^2. \end{aligned} \tag{8}$$

The last term $E^S(\cdot,\cdot)$ (smoothness) enforces saliency consistency between adjacent superpixels, which is defined as:

$$E^S(x_i, x_j) = w_{i,j}(S_{n,t}(x_i) - S_{n,t}(x_j))^2 \tag{9}$$

where the weight $w_{i,j} = e^{-\|f_c(x_i) - f_c(x_j)\|_1}$ defines a similarity measure for adjacent superpixels. Note that, $f_c(x_i)$ indicates the color feature of superpixel x_i. The parameters αs are the positive coefficients for balancing the relative influence between various terms and we set $\alpha_1 = \alpha_2 = \alpha_3 = 0.2$. Equ. 7 can be solved by convex optimization.

Since t-th frame's saliency $\{S_{n,t}\}_{t=1}^N$ is obtained for each V^n, we can establish two color histograms $\{H_f^t, H_b^t\}$ for foreground and background. These two histograms are computed as the weighted aggregation of superpixels color values, where the weights for superpixels x are $S_{n,t}(x)$ and $(1-S_{n,t}(x))$, respectively. We then use this appearance model to give an initial saliency estimation $S_{n,t}^I$ for next $\{t+1\}$-th frames $\{I_{t+1}^n\}_{n=1}^N$. Thus, we can obtain salient area $\mathcal{F}_{t+1}^n$ and background area $\mathcal{B}_{t+1}^n$. Finally, we use the saliency cues and the saliency integration method mentioned above to produce the final co-saliency for $\{t+1\}$-th frames $\{I_{t+1}^n\}_{n=1}^N$. For the first frame (t=1), the initial saliency estimation is generated by an average of various saliency methods. When t>1, the initial saliency is obtained through appearance model $\{H_f^{t-1}, H_b^{t-1}\}$ of the prior frame, which offers a more correct estimation result. After we obtain the co-saliency results for all the videos, we repeat the detection process in the reverse order for improved results. The obtained appearance model $\{H_f^t, H_b^t\}_{t=1}^N$ offers initial saliency estimation for prior frame I_{t-1}^n in turn, three kinds of saliency cues are then computed and further combined through Eqn. 7 for generating final saliency results.

B. Video Co-segmentation

Our co-saliency guided co-segmentation method extracts the common salient regions by minimizing an energy function. The co-segment energy is derived from a conditional random field (CRF) model that incorporates our aforementioned co-saliency measure. Given a video V^n and the set of superpixels $\mathcal{X}^n = \{X_t^n\}_t = \{x_{t,i}^n\}_{t,i}$, our task is to find $\mathcal{L}^n = \{l_{t,i}^n\}_{t,i}$ so that $l_{t,i}^n \in \{0,1\}$ indicates whether superpixel $x_{t,i}^n \in X_t^n$ belongs to the common salient object or the background. We cast this problem into a spatiotemporal energy minimization framework. For video V^n, the optimal binary labeling is obtained by minimizing the following energy function over the labels $\mathcal{L}^n$:

$$\begin{aligned} \mathcal{F}(\mathcal{L}^n) = & \sum_{t,i} \mathcal{S}(l_{t,i}^n) + \sum_{t,i} \mathcal{A}(l_{t,i}^n) + \sum_{t,i} \mathcal{G}(l_{t,i}^n) \\ & + \sum_t \sum_{i,j \in \aleph_s} \mathcal{V}(l_{t,i}^n, l_{t,j}^n) + \sum_t \sum_{i,j \in \aleph_t} \mathcal{W}(l_{t,i}^n, l_{t,j}^{n+1}), \end{aligned} \tag{10}$$

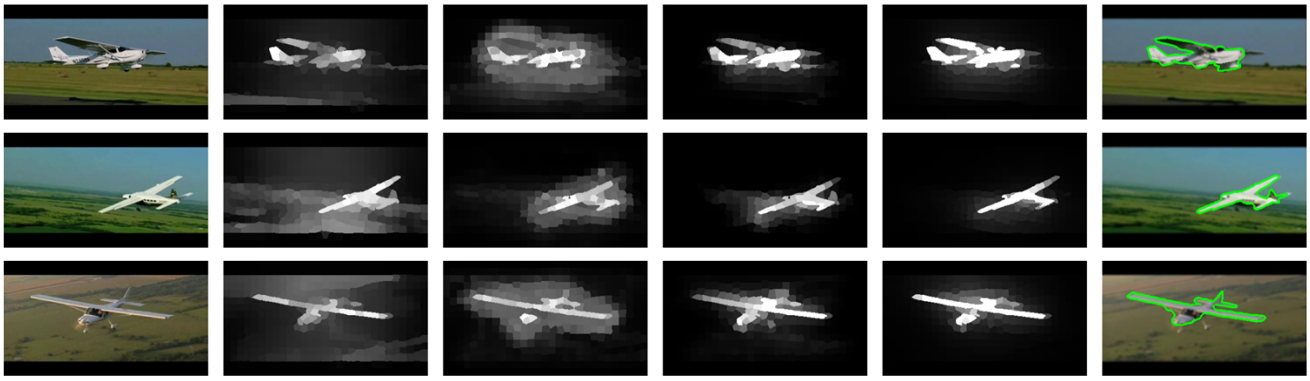


Fig. 2. Illustration of our co-saliency based co-segmentation approach. (a) Input video group $\mathcal{V}$. (b) Initial saliency estimation $S^I_{n,1}$ by averaging the outputs of various saliency methods [1], [2], [3]. (c) Intra-video appearance estimation. (d) Inter-video appearance estimation. (e) Video co-saliency results. (f) Our video co-segmentation results.

where $\mathcal{S}(\cdot)$ is the co-saliency unary term, $\mathcal{A}(\cdot)$ is the intra-video appearance term, $\mathcal{G}(\cdot)$ denotes the inter-video appearance term, $\mathcal{V}(\cdot,\cdot)$ and $\mathcal{W}(\cdot,\cdot)$ are the pairwise potentials representing the spatial smoothness term and temporal smoothness term, respectively. $i,j \in \aleph_s$ indicates that two superpixels $x^n_{t,i}, x^n_{t,j}$ are spatially connected (they are in the same frame and are adjacent). Similarly, $i,j \in \aleph_t$ represents two superpixels $x^n_{t,i}, x^{n+1}_{t,j}$ are temporally connected. Two superpixels are temporally connected if they are in subsequent frames and connected via optical flow.

1) Co-saliency term: The video co-saliency captures the likelihood of a superpixel belonging to the foreground, therefore, we define the co-saliency term $\mathcal{S}$ for superpixel $x^n_{t,i}$ as:

$$\mathcal{S}(l^n_{t,i}) = -l^n_{t,i} \log S(x^n_{t,i}) - (1 - l^n_{t,i}) \log(1 - S(x^n_{t,i})). \quad (11)$$

This data term relies on our discriminative co-saliency maps estimated by the prior step. Our co-saliency maps are estimated across multiple videos, and thus are video independent.

2) Intra-video appearance term: Since a co-saliency score also indicates spatiotemporally salient objects in a video, we establish the foreground/background appearance models for each video too. Given video V^n, we construct two color histograms, one for the foreground and the other for the background. Each superpixel in video V^n is stacked into histograms and weighted by its foreground likelihood according to saliency scores $\{S^n_t\}_t$. We define the intra-video appearance term $\mathcal{A}$ for superpixel $x^n_{t,i}$ as:

$$\mathcal{A}(l^n_{t,i}) = -l^n_{t,i} \log p^A(x^n_{t,i}) - (1 - l^n_{t,i}) \log(1 - p^A(x^n_{t,i})), \quad (12)$$

where $p^A(\cdot)$ is the likelihood that the superpixel belongs to the foreground given by the intra-video appearance model.

3) Inter-video appearance term: We employ an inter-video appearance term based on the global appearance model $\{H^t_f, H^t_b\}_t$ generated in the previous co-saliency step. Let $p^G(x^n_{t,i})$ be the likelihood of superpixel $x^n_{t,i}$ for the foreground according to the global appearance model, we define the inter-video appearance term $\mathcal{G}$ for superpixel $x^n_{t,i}$ as:

$$\mathcal{G}(l^n_{t,i}) = -l^n_{t,i} \log p^G(x^n_{t,i}) - (1 - l^n_{t,i}) \log(1 - p^G(x^n_{t,i})). \quad (13)$$

A superpixel that has a similar color histogram to the foreground (or background) gets a higher cost if labeled otherwise. Based on the global appearance model, we process videos one by one, making the run time of our method linear with the number of videos.

4) Spatial and temporal smoothness terms: The spatial smoothness term $\mathcal{V}$ and the temporal smoothness term $\mathcal{W}$ encourage label smoothness over the spatially and temporally neighboring superpixels. We use a contrast-dependent function defined in [23], [25], which favors assigning the same label to neighboring superpixels that have similar colors.

The spatial consistency term $\mathcal{V}$ computed between spatially adjacent pixels $x^n_{t,i}$ and $x^n_{t,j}$ is defined as

$$\mathcal{V}(l^n_{t,i}, l^n_{t,j}) = \delta(l^n_{t,i}, l^n_{t,j}) e^{-\|f_c(x^n_{t,i}) - f_c(x^n_{t,j})\|^2_2}, \quad (14)$$

where $f_c(x^n_{t,i})$ indicates the color feature of superpixel $x^n_{t,i}$ and $\delta(\cdot)$ denotes the Dirac delta function, which is 0 when $l^n_{t,i} \neq l^n_{t,j}$.

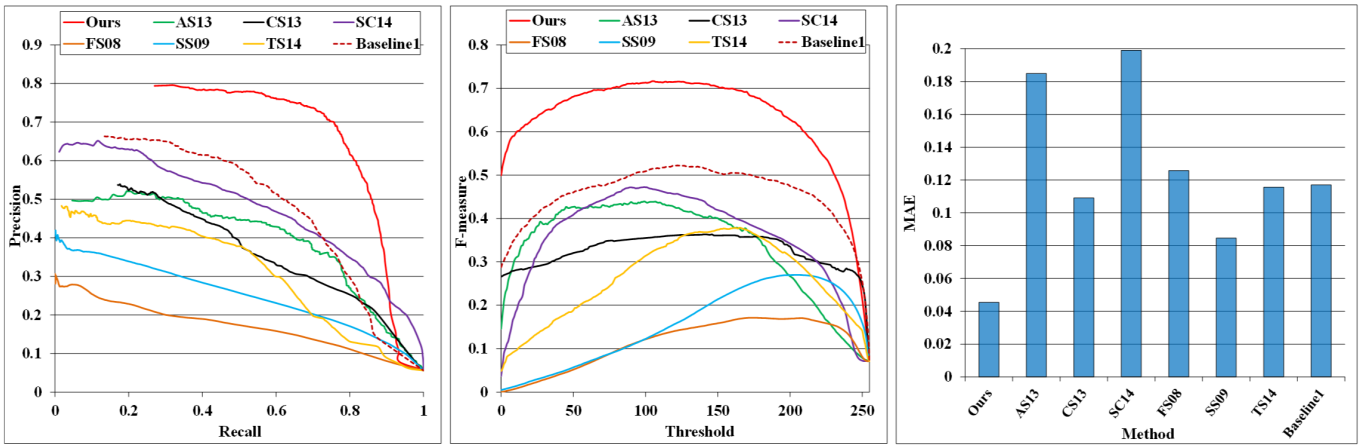
Similarly, the temporal consistency term $\mathcal{W}$ is defined as

$$\mathcal{W}(l^n_{t,i}, l^{n+1}_{t,j}) = \delta(l^n_{t,i}, l^{n+1}_{t,j}) e^{-\|f_c(x^n_{t,i}) - f_c(x^{n+1}_{t,j})\|^2_2}. \quad (15)$$

In Equ. 10, co-saliency term $\mathcal{S}(\cdot)$, intra-video appearance term $\mathcal{A}(\cdot)$ and inter-video appearance term $\mathcal{V}(\cdot,\cdot)$, and spatial smoothness $\mathcal{V}(\cdot,\cdot)$ and temporal consistency $\mathcal{W}(\cdot,\cdot)$ terms are efficiently combined for object co-segmentation, which can be effectively minimized via graph cuts [52].

IV. Experimental Results

Our approach automatically highlights and co-segments the common salient regions from a group of video sequences using the video co-saliency cue. We present our co-saliency results and co-segmentation on two datasets, including the Safari dataset [21] and a new video object co-segmentation dataset collected by ourselves. The Safari dataset [21] was proposed for video co-segmentation, which contains 5 classes of animals and a total of 9 videos. 5 frames of each video have pixel-level annotations for the object labels. Up to now there is no previous work for the video co-saliency issue. We prepared a large benchmark dataset, which has 10 different



(a) (b) (c)

Fig. 3. Statistical comparisons on the Safari dataset [21] between our method and other approaches: AS13 [3], CS13 [14], SC14 [15], FS08 [7], SS09 [8], TS14 [9], as well as *Baseline1*. From left to right: (a) precision recall curve, (b) F-score, (c) average MAE.

Fig. 4. Visual comparison of previous approaches: AS13 [3], CS13 [14], SC14 [15], FS08 [7], SS09 [8], TS14 [9] with our method using two video groups from the Safari dataset [21]. Our method consistently produces video co-saliency maps most similar to the ground truth.

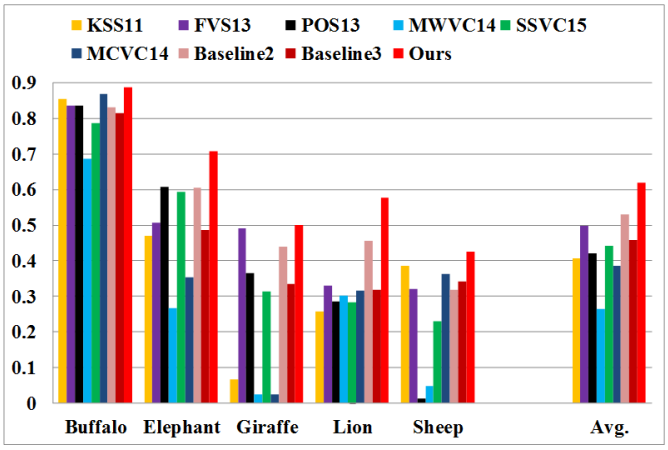


Fig. 5. Comparisons of average Jaccard scores on the Safari dataset [21] between our method and other approaches: KSS11 [23], FVS13 [25], POS13 [24], MWVC14 [21], SSVC15 [22], and MCVC14 [19].

video groups including 38 videos in total, for benefiting future research. Each video group consists of multiple relevant videos with similar appearances of the foreground elements, which are suitable for co-saliency and co-segmentation problems.

We compare our method to six state-of-the-art saliency methods including absorbing Markov chain based AS13 [3] for image saliency, and two image co-saliency methods: CS13 [14], SC14 [15], and video saliency methods: FS08 [7], SS09 [8], and TS14 [9]. For evaluating our co-segmentation performance, we further provide both qualitative as well as quantitative comparison with several competitors including three video segmentation works: key-segments for video segmentation KSS11 [23], fast video segmentation FVS13 [25], and primary object regions for video segmentation POS13 [24] and three video co-segmentation algorithms: maximum weight cliques based video co-segmentation MWVC14 [21], multi-class video co-segmentation MCVC14 [19], and SIFT-flow based robust video co-segmentation SSVC15 [22].

A. Results on Safari Dataset

1) Video co-saliency results: For evaluating the performance of our co-saliency model, we provide both qualitative as well as quantitative comparison with several state-of-the-art methods: AS13 [3], CS13 [14], SC14 [15], FS08 [7], SS09 [8], and TS14 [9]. For demonstrating the benefits of our inter-video saliency cue, which explores the relationship between common salient regions, we further offer a baseline *Baseline1*. *Baseline1* indicates our saliency results without inter-video correspondences, which are estimated via foreground/background histograms from each video independently instead of estimating $\{H_t^f, H_t^b\}$ across videos in Equ. 3.

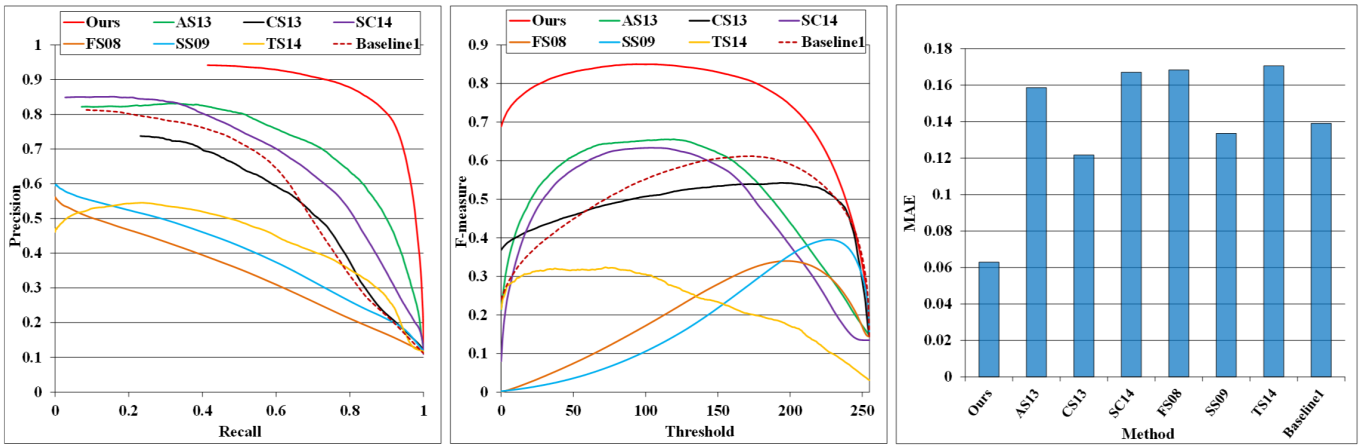


Fig. 7. Statistical comparisons on our dataset between our method and other approaches: AS13 [3], CS13 [14], SC14 [15], FS08 [7], SS09 [8], TS14 [9], as well as *Baseline1*. From left to right: (a) precision recall curve, (b) F-score, (c) average MAE.

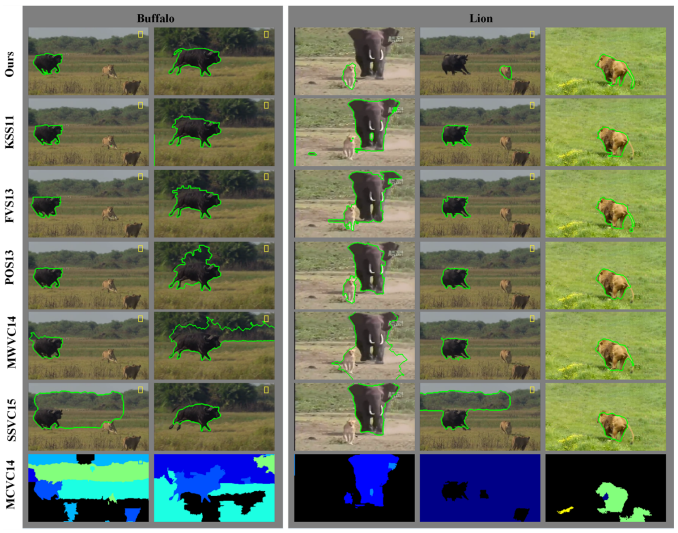


Fig. 6. Co-segmentation results on the Safari dataset [21].

Fig. 8. Visual comparison of previous approaches: AS13 [3], CS13 [14], SC14 [15], FS08 [7], SS09 [8], TS14 [9] with our method using three video groups from our dataset. Our method consistently produces video co-saliency maps most similar to the ground truth.

In quantitative evaluation, we use three widely used criteria: PR (precision-recall) curve, F-score and MAE (mean absolute errors). The PR curves of our algorithm are shown in Fig. 3(a), which indicates that our method achieves the best performance. The F-score is depicted in Fig. 3(b), where our method also performs better than the others. MAE scores are presented in Fig. 3(c). As shown, our method successfully reduces the error by **48.30%** over the second best algorithm (SC14), and **53.85%** over the third best algorithm (SS09). Furthermore, the significant improvement over *Baseline1* indicates the benefit of considering all videos in a joint manner. Fig. 4 shows a visual comparison of different methods over two video groups from the Safari dataset. Consistent with previous quantitative reports, our method generates more reasonable maps compared with other approaches.

2) Video co-segmentation results: Our framework simultaneously segments common salient regions from a group of related videos. In this subsection, we compare our method with three top performing video segmentation methods: KSS11 [23], FVS13 [25], and POS13 [24] as well as three state-of-the-art video co-segmentation methods: MWVC14 [21], MCVC14 [19], and SSVC15 [22], and show quantitative results on the Safari database. As [19] is designed to return multi-class segments, we report results for the segment best matching the ground-truth. We also present two baselines: *Baseline2* and *Baseline3* for further evaluating the improvement of our CRF-based co-segmentation strategy and the effect of our co-saliency cue,. *Baseline2* is obtained by binarizing the co-saliency map using threshold (0.5), and

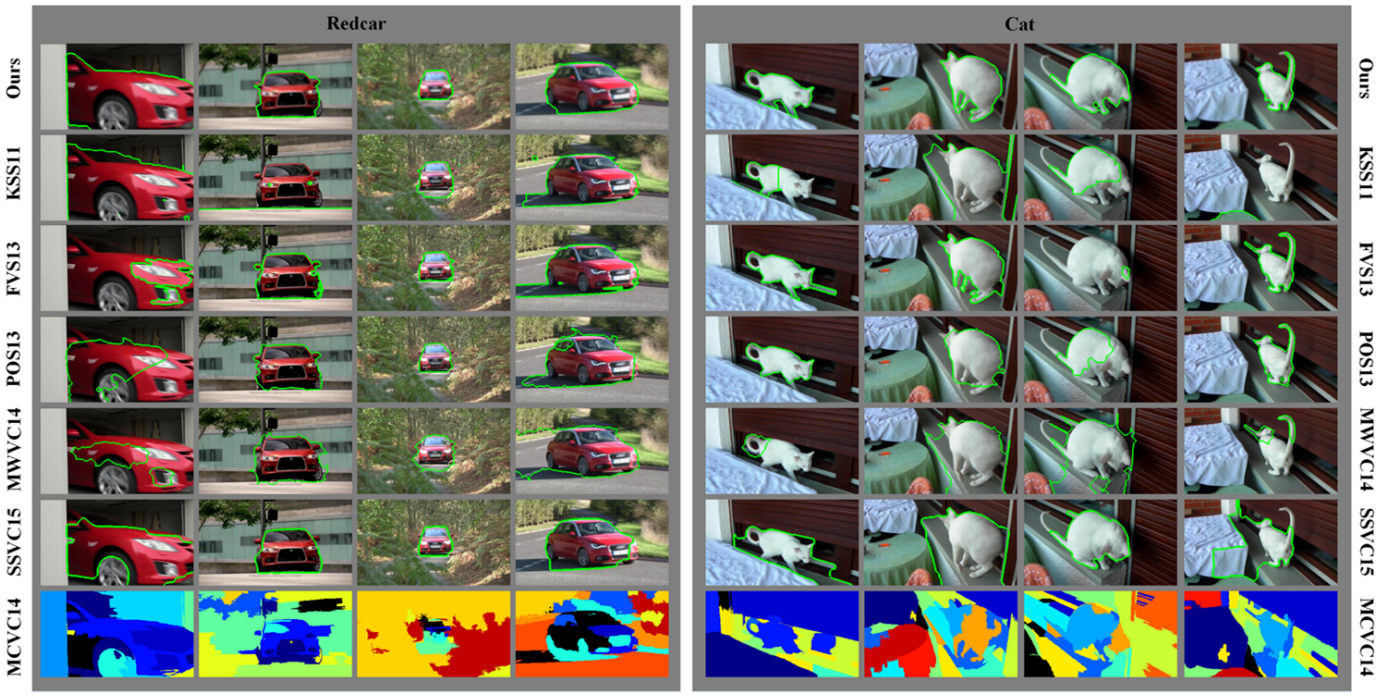


Fig. 10. Co-segmentation results on our benchmark dataset.

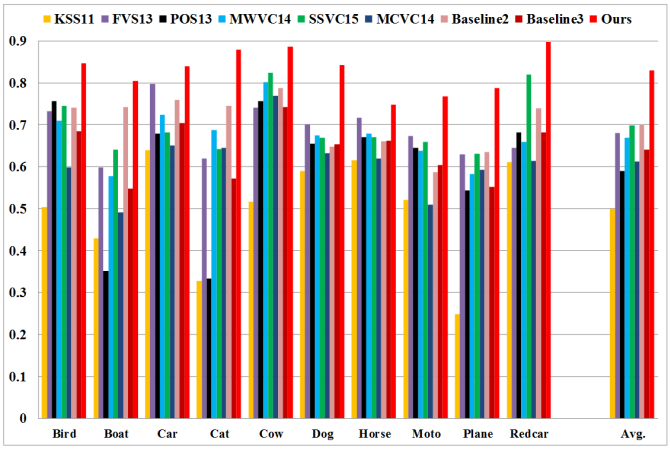


Fig. 9. Comparisons of average Jaccard scores on our dataset between our method and other approaches: KSS11 [23], FVS13 [25], POS13 [24], MWVC14 [21], SSVC15 [22], and MCVC14 [19].

Baseline3 is computed through Equ. 10 without the co-saliency term and the inter-video appearance term.

To evaluate the performance on video co-segmentation, we measure the Jaccard score as the intersection over union of the result and ground truth segmentations. As depicted in Fig. 5, the proposed method improves on the state of the art. One reason for the superior performance compared with video segmentation methods KSS11 [23], FVS13 [25], and POS13 [24] is that our method fully explores the correspondence between similar objects from different videos. A visual comparison is demonstrated in Fig 6. Our method yields better co-segmentation performances than the video co-segmentation method, MWVC14 [21], MCVC14 [19], and SSVC15 [22], thanks to fact that our co-saliency prior integrates various object characteristics that are indicative for foregrounds and effective saliency assumptions such as contrast and location bias.

B. Results on Our Dataset

1) Video co-saliency results: We also compared our results with the same six state-of-the-art methods: AS13 [3], CS13 [14], SC14 [15], FS08 [7], SS09 [8], and TS14 [9] as in Sec. IV-A. Fig. 7 presents this comparison, showing again that our method performs better than the state-of-the-art methods in terms of PR curves. And it has a wider range of high F-measure compared to others. Furthermore, the lowest MAE of our method displayed in Fig. 7 (c) indicates the similarity between our saliency maps and the ground truth. Another interesting aspect of the proposed our method method is how saliency results are supported by exploiting commonality of objects in a video collection via a global appearance model. Without this global model, significant increase of MAE can be observed when we compare *Baseline1* and our method. Fig. 8 shows the visual comparison by different methods. It is obvious that the co-saliency maps generated by our algorithm highlight very accurately the salient regions with few noisy areas.

2) Video co-segmentation results: The proposed our method is compared with five previously proposed methods: KSS11 [23], FVS13 [25], POS13 [24], MWVC14 [21], MCVC14 [19], and SSVC15 [22] and two baselines: *Baseline2* and *Baseline3* on the collected dataset. As presented in Fig. 9, our method can achieve highest average Jaccard scores, significantly outperforming other methods. Better performance compared with two baselines verifies the contribution of the

TABLE I
COMPARISON OF AVERAGE RUN TIME (SECONDS PER FRAME) ON SAFARI DATASET

Method	our method	MWVC14 [21]	MCVC14 [19]	SSVC15 [22]
Time(s)	2.24+1.62	8.9	25.6	16.7

proposed co-saliency cue and the effectiveness of the co-segmentation energy function. Qualitative examples of our method are depicted in Fig. 10. Results in the *Redcar* group exhibit that our approach can co-segment the objects with large shape deformations or various motion patterns. From the *Cat* group, we can observe that our method is applicable for difficult scenarios where there is significant color similarity between foreground and background since our method utilizes motion information and location cues too. The qualitative results clearly demonstrate that our algorithm achieves significantly better performance than the state-of-the-art.

C. Runtime Analysis

We carry out time consumption analysis of the proposed method on a personal computer equipped with Intel i5 CPU of 2.50 GHz and 4GB RAM. With our unoptimized MATLAB code, our method takes around 3.8s per frame on Safari dataset [21]. The computational cost of video co-saliency stage takes 2.2s, while our video co-saliency stage takes around 1.6s. The average run time of currently top-performing video co-segmentation methods: MWVC14 [21], MCVC14 [19], and SSVC15 [22] are presented in Table 1. The run time excludes optical flow computation, which all methods require as input. As shown in Table 1, our run time is much faster than other methods.

V. CONCLUSION

The presented method simultaneously partitions a group of videos depicting the same or similar object into foreground and background. Our method estimates the foreground and background appearance distributions from several videos using a novel joint video co-saliency concept. Our method adopts various discriminative saliency cues to find the common salient regions. It is based on the observation that common foregrounds share similar appearances across videos and salient areas highly contrast with surrounding backgrounds. We have proposed a location saliency cue based on the saliency of a prior frame, which is more accurate than traditional center-bias assumption. We also proposed a video co-segmentation energy function considering a co-saliency prior and a joint foreground appearance model. Experimental results have validated the superiority of our approach. Compared to former state-of-the-art methods, our method accurately locates complete salient objects with complex motion patterns, even in the presence of a cluttered background.

REFERENCES

[1] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proceedings of ECCV*, 2012.

[2] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings of IEEE CVPR*, 2012.

[3] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *Proceedings of IEEE ICCV*, 2013.

[4] Z. Ren, S. Gao, L.-T. Chia, and I. W. H. Tsang, "Region-Based Saliency Detection and Its Application in Object Recognition," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 769–779, 2014.

[5] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. on Image Processing*, vol. 25, no. 11, pp. 5025-5034, 2016.

[6] W. Wang and J. Shen, Y. Yu, and K-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Trans. on Visualization and Computer Graphics*, in press, doi://10.1109/TVCG.2016.2600594, 2016.

[7] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proceedings of IEEE CVPR*, 2008.

[8] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of vision*, 2009.

[9] F. Zhou, S. B. Kang, and M. Cohen, "Time-mapping using space-time saliency," in *Proceedings of IEEE CVPR*, 2014.

[10] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-Based Spatiotemporal Saliency Detection," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 9, pp. 1522–1540, 2014.

[11] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. on Image Processing*, vol. 24, no. 11, pp. 4185-4196, 2015.

[12] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. on Image Processing*, 2011.

[13] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proceedings of IEEE CVPR*, 2011.

[14] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. on Image Processing*, 2013.

[15] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. on Image Processing*, 2014.

[16] D. Zhang, J. Han, and L. Shao, "Co-saliency Detection Based on Intrasaliency Prior Transfer and Deep Intersaliency Mining," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1163-1176, 2016.

[17] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of Co-salient Objects by Looking Deep and Wide," *International Journal of Computer Vision*, vol. 120, pp. 215-232, 2016.

[18] J. Guo, Z. Li, L.-F. Cheong, and S. Z. Zhou, "Video co-segmentation for meaningful action extraction," in *Proceedings of IEEE ICCV*, 2013.

[19] W.-C. Chiu and M. Fritz, "Multi-class video co-segmentation with a generative multi-video model," in *Proceedings of IEEE CVPR*, 2013.

[20] H. Fu, D. Xu, B. Zhang, and S. Lin, "Object-based multiple foreground video co-segmentation," in *Proceedings of IEEE CVPR*, 2014.

[21] D. Zhang, O. Javed, and M. Shah, "Video object co-segmentation by regulated maximum weight cliques," in *Proceedings of ECCV*, 2014.

[22] W. Wang, J. Shen, X. Li, and F. Porikli, "Robust video object cosegmentation," *IEEE Trans. on Image Processing*, vol. 24, no. 10, pp. 3137–3148, 2015.

[23] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proceedings of IEEE ICCV*, 2011.

[24] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *Proceedings of IEEE CVPR*, 2013.

[25] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proceedings of IEEE ICCV*, 2013.

[26] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2011.

[27] J. Yang and M. Yang, "Top-down visual saliency via joint CRF and dictionary learning," in *Proceedings of IEEE CVPR*, 2012.

[28] D. Gao, M. Vijay, and V. Nuno, "The discriminant center-surround hypothesis for bottom-up saliency," *Proceedings of NIPS*, pp. 497–504, 2008.

[29] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of IEEE CVPR*, 2014.

[30] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2010.

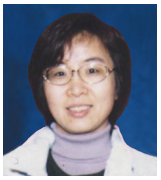
[31] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proceedings of ECCV*, 2010.

[32] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proceedings of CVPR*, 2015.

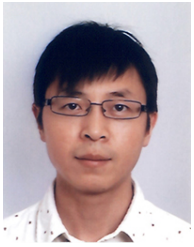
[33] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, in press, doi://10.1109/TPAMI.2017.2662005, 2017.
[34] D. Zhang, H. Fu, J. Han, and F. Wu, "A Review of Co-saliency Detection Technique: Fundamentals, Applications, and Challenges," in *Proceedings of CVPR*, 2016.
[35] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proceedings of ECCV*, 2010.
[36] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, "Track to the future: Spatio-temporal video segmentation with long-range motion cues," in *Proceedings of IEEE CVPR*, 2011.
[37] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proceedings of IEEE CVPR*, 2012.
[38] W. T. Li, H. S. Chang, K. C. Lien, H. T. Chang, and Y. C. F. Wang, "Exploring visual and motion saliency for automatic video object extraction," *IEEE Trans. on Image Processing*, 2013.
[39] W. Brendel and S. Todorovic, "Video object segmentation by tracking regions," in *Proceedings of IEEE ICCV*, 2009.
[40] C. Xu, C. Xiong, and J. J. Corso, "Streaming hierarchical video segmentation," in *Proceedings of ECCV*, 2012.
[41] T. Ma and L. J. Latecki, "Maximum weight cliques with mutex constraints for video object segmentation," in *Proceedings of IEEE CVPR*, 2012.
[42] J. Yang, G. Zhao, J. Yuan, X. Shen, Z. Lin, B. Price, and J. Brandt, "Discovering primary objects in videos by saliency fusion and iterative appearance estimation," *IEEE Trans. on Circuits and Systems for Video Technology*, 2016.
[43] J. C. Rubio, J. Serrat, and A. López, "Video co-segmentation," in *Proceedings of IEEE ACCV*, 2012.
[44] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, and R. Sukthankar, "Weakly supervised learning of object segmentations from web-scale video," in *Proceedings of ECCV*, 2012.
[45] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, "Discriminative segment annotation in weakly labeled video," in *Proceedings of CVPR*, 2013.
[46] X. Liu, D. Tao, M. Song, Y. Ruan, C. Chen, and J. Bu, "Weakly supervised multiclass video segmentation," in *Proceedings of CVPR*, 2014.
[47] W. Wang, and J. Shen, "Higher-order image co-segmentation," *IEEE Trans. on Multimedia*, 2016.
[48] Y. J. Lee, and K. Grauman, "Foreground focus: Finding meaningful features in unlabeled images," in *Proceedings of BMVC*, 2008.
[49] A. Chandrashekar, L. Torresani, R. Granger, "Learning what is where from unlabeled images: joint localization and clustering of foreground objects," *Machine learning*, 2014.
[50] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2012.
[51] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, 2011.
[52] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2001.

Wenguan Wang received the B.S. degree in computer science and technology from the Beijing Institute of Technology in 2013. He is currently working toward the Ph.D. degree in the School of Computer Science, Beijing Institute of Technology, Beijing, China. His current research interests include computer vision and deep learning.

Jianbing Shen (M'11-SM'12) is a Professor with the School of Computer Science, Beijing Institute of Technology, Beijing, China. His research interests include computer vision and multimedia processing. He has published about 60 journal and conference papers such as *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Image Processing*. He has also obtained many flagship honors including the Fok Ying Tung Education Foundation from Ministry of Education, the Program for Beijing Excellent Youth Talents from Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from Ministry of Education. His research interests include computer vision and multimedia processing. He is on the editorial boards of *Neurocomputing*.



Hanqiu Sun is an Associate Professor with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Shatin, Hong Kong. She received the Ph.D. degree in computer science from University of Alberta, Alberta, ON, Canada. Her research interests include virtual reality, interactive graphics, image/video editing, and touch-enhanced simulations.



Ling Shao (M'09-SM'10) is a Professor with the School of Computing Sciences at the University of East Anglia, Norwich, UK. His current research interests include computer vision, image processing, pattern recognition, and machine learning. He is an Associate Editor of *IEEE Transactions on Image Processing*, *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Transactions on Neural Networks and Learning Systems*, and other journals.