Robust Video Object Cosegmentation

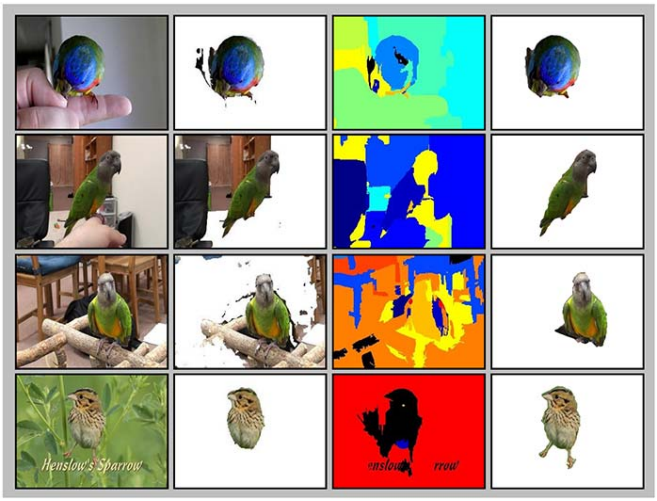
Wenguan Wang, Jianbing Shen, *Senior Member, IEEE*, Xuelong Li, *Fellow, IEEE*, and Fatih Porikli, *Fellow, IEEE*

***Abstract*— With ever-increasing volumes of video data, automatic extraction of salient object regions became even more significant for visual analytic solutions. This surge has also opened up opportunities for taking advantage of collective cues encapsulated in multiple videos in a cooperative manner. However, it also brings up major challenges, such as handling of drastic appearance, motion pattern, and pose variations, of foreground objects as well as indiscriminate backgrounds. Here, we present a cosegmentation framework to discover and segment out common object regions across multiple frames and multiple videos in a joint fashion. We incorporate three types of cues, i.e., intraframe saliency, interframe consistency, and across-video similarity into an energy optimization framework that does not make restrictive assumptions on foreground appearance and motion model, and does not require objects to be visible in all frames. We also introduce a spatio-temporal scale-invariant feature transform (SIFT) flow descriptor to integrate across-video correspondence from the conventional SIFT-flow into interframe motion flow from optical flow. This novel spatio-temporal SIFT flow generates reliable estimations of common foregrounds over the entire video data set. Experimental results show that our method outperforms the state-of-the-art on a new extensive data set (ViCoSeg).**

***Index Terms*— Video object co-segmentation, energy optimization, object refinement, spatio-temporal scale-invariant feature transform (SIFT) flow.**

I. Introduction

With the faster growth of video data, efficient and automatic extraction of the interest object from multiple videos is quite important and very challenging. Maybe these objects of interest exhibit drastically different in their appearance or motions. Moreover, foreground appearance or motions from various videos are much different, while possibly low contrast with the background. These challenges cause great difficulties on existing video segmentation techniques [3], [11], [12], [16], [19], [33], which usually benefit from visual cues such as motion or appearance. Additionally, these methods rely on the assumption that the motion or appearance of object is dramatically distinct from background, which is against the situation as we mentioned before. Moreover, the lack of taking into account the joint information between videos leads to unsatisfactory performance of these methods designed for single video on this issue (see Fig. 1 (b)).



(a) (b) (c) (d)

Fig. 1. Video co-segmentation. (a) Input videos where objects have large variations. (b) Results by [32], which lacks the joint information between the videos. (c) Results by video co-segmentation method of [25]. Over-fragmentation is visible. Also, parts of foregrounds (e.g. bird) are merged into background as its global model heavily relies on the chroma and motion. (d) Our video object co-segmentation results.

In contrast to previous object segmentation methods for a single video, video co-segmentation has been proposed to extract the main common object from a set of related videos. Video co-segmentation utilizes visual properties across multiple videos to infer the object of interest with the absence of priori information about videos or foregrounds. There are a few methods designed for this problem till now [21], [22], [25]. While these approaches make quite strong assumptions on the motion patterns or appearance of foreground. For example, Rubio *et al.* [21] make assumptions that the foreground objects from different videos have similar motion patterns and similar appearance model which is

Manuscript received July 28, 2014; revised December 13, 2014 and March 13, 2015; accepted May 21, 2015. Date of publication June 1, 2015; date of current version June 12, 2015. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2013CB328805, in part by the National Natural Science Foundation of China under Grant 61272359 and Grant 61125106, in part by the Australian Research Council through the Discovery Projects Funding Scheme under Grant DP150104645, in part by the Key Research Program through the Chinese Academy of Sciences under Grant KGZDEW-T03, and in part by the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Andrea Cavallaro. *(Corresponding author: Jianbing Shen.)*

W. Wang and J. Shen are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wangwenguan@bit.edu.cn; shenjianbing@bit.edu.cn).

X. Li is with the Center for OPTical IMagery Analysis and Learning, School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: xuelong_li@nwpu.edu.cn).

F. Porikli is with the Research School of Engineering, Australian National University, Canberra, ACT 0200, Australia, and also with the NICTA, Eveleigh, NSW 2015, Australia (e-mail: fatih.porikli@anu.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2438550

distinct from the background. Chen *et al.* [22] emphasize that the coherent motion of regions and similar appearance are able to conduct the segmentation. Additionally, one general limitation of these approaches [21], [22] is that the set of videos is assumed to be similar or related for foregrounds and backgrounds. Chiu and Fritz [25] treat the task of video co-segmentation as a multi-class labeling problem, but its classification results heavily rely on the chroma and motion features (see Fig. 1 (c)). Totally, these previous video co-segmentation approaches [21], [22], [25] have two main limitations. First, both approaches abuse motion or appearance based cues and ignore the fact that there are considerable videos with the common object low contrast with the background. Second, in both approaches, the process of inferring common objects does not effectively explore the correspondence of objects from different videos, which is essential for the task of video co-segmentation. These methods simply assume that the objects are similar in motion patterns or appearance, which is not suitable for the scene that includes objects with large variations in appearance or motion. Besides, there are considerable videos that include some frames not containing the common object of the whole video sequence. For instance, the foreground object moves out of camera or the switching between video shots. However, this general fact is ignored by most previous work in both video object segmentation and co-segmentation methods. Most of methods assume that the foreground object appears in every frame, and hence they are unable to perform well for this issue.

This paper presents a co-segmentation framework for detecting and segmenting out common objects from multiple, contextually related videos without imposing above constraints. In our approach, we explore the underlying properties of video objects in three levels: intra-frame saliency, inter-frame consistency and across-video correspondence. Based on these properties, we introduce a spatio-temporal SIFT flow descriptor to capture the relationship between foreground objects. We establish an object discovery energy function utilizing the spatio-temporal SIFT flow and inter-frame consistency to discover the common objects. Our source code will be publicly available online.[1]

Compared to existing video co-segmentation approaches, the proposed method offers following contributions:

- A novel video co-segmentation method is proposed for automatically segmenting out the foreground object with low constraint for their appearance and motion patterns.
- We are the first to fully explore the properties of foreground object in video: intra-frame saliency, inter-frame consistency and across-video similarity. These important cues are further formulated into our video co-segmentation framework as the optimization problems.
- An efficient spatio-temporal SIFT flow is developed to build reliable correspondences between different videos, which can infer the common object over entire video dataset and refine the segmentation accuracy for objects.
- We are the first to emphasize the fact that some frames perhaps do not contain the common object. A novel object discovery energy function is proposed to discover the common object with this situation by utilizing the proposed spatio-temporal SIFT flow and those properties of foreground object.

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

II. Related Work

We give a short overview of the previous work along two major themes: video co-segmentation and video object segmentation techniques below.

Video Co-Segmentation: Video co-segmentation has received attentions only recently, thus there are very few methods [21], [22], [25] specially designed for this purpose to the best of our knowledge.

Rubio *et al.* [21] provided an iterative optimization framework to achieve such a video co-segmentation task. This work is based on a dense feature matching process executed on region and tube levels using joint appearance and motion models of the foreground and background. While this approach made quite strong assumptions that foreground objects from different videos have similar motion patterns and similar appearance models which are distinct from background. Obviously, its applicability is limited by its unmatched assumptions. The work by Chen *et al.* [22] utilized the motion coherence and appearance cues to separate the common object in a pair of related videos. However, this method attempted to group the regions into foreground and background according to the coherent motion and similar appearance, which leads to unsatisfactory performance for the video with similar foreground and background motions or appearance. Moreover, both Rubio *et al.* [21] and Chen *et al.* [22] required the input videos to be similar. Therefore, they may fail for cases that have large variations in foreground appearance and complex backgrounds. Chiu and Fritz [25] performed multi-class video co-segmentation by building a non-parametric Bayesian model based on Dirichlet Processes that relies on the chroma similarity and motion distinction constraints. As a result, the discrimination power of this model is limited in complex scenarios. When the input videos with more common scenario, their results sometimes are consistent with the regions that exhibit coherent appearance or motion instead of a particular object. It can be seen that video object co-segmentation is still an emerging research problem to be intensively investigated.

Video Object Segmentation: The goal of video object segmentation is to detect the primary object and extract the object from a single video. There has been a large body of work concentrating on this task last decade. Video object segmentation methods can be broadly classified into two categories: interactive (supervised) methods and automatic (unsupervised) methods. For interactive video object segmentation [2], [4], [5], [11]–[13], [28], user interactions and optimization techniques employing motion and appearance constraints are often introduced to produce high quality segmentation results.

Our method is more closed to unsupervised video object segmentation. Unsupervised video object segmentation aims at autonomously merging pixels into foreground or

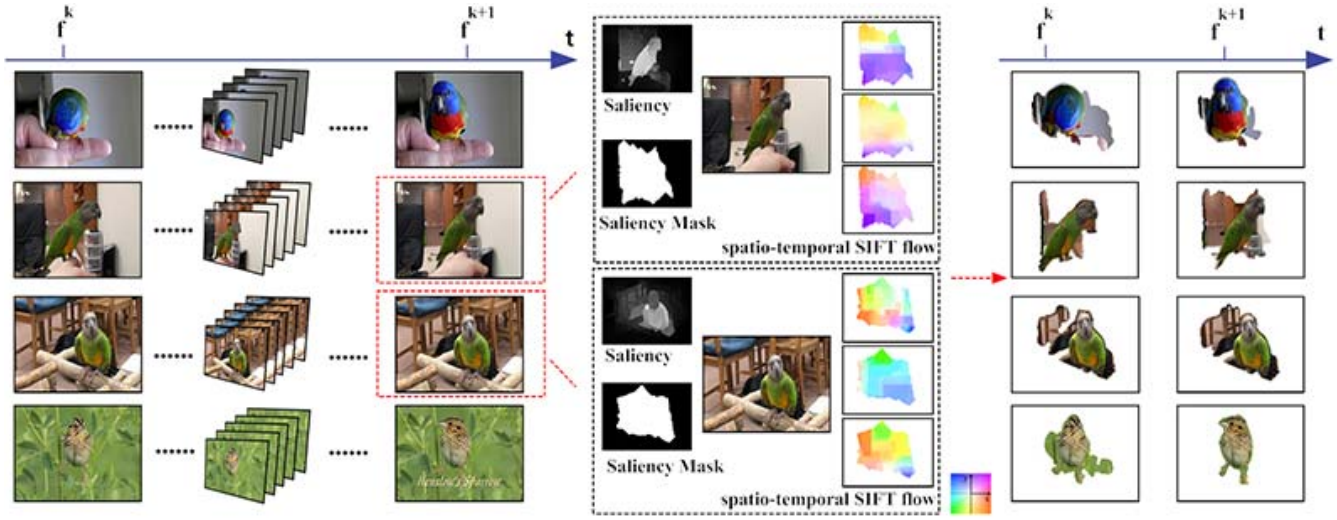


Fig. 2. Overview of our object discovery step. (a) Four input videos where bird is the common object. This object discovery process does not need to be performed at full frame rate. There are five frames between frame f^k and frame f^{k+1}. (b) Saliency information and spatio-temporal SIFT flow are introduced into this step to get the common object in video set. (c) Output of the object discovery step is a coarse estimation for the common object regions in each frame based on the object discovery energy function as in (13).

background within their video. Earlier automatic segmentation methods [6], [14], [15], [30], [31] employed appearance or motion based cues for a bottom-up segmentation. Several methods [16], [19], [32] were proposed to select primary object regions in object proposal domain based on the notion of what a generic object looks like. These methods benefit from the work of object hypotheses proposals [8]–[10] that offer considerable object candidates in every image/frame. Therefore, segmenting video object is transformed into an object region selection problem. In this selection process, both motion and appearance cues are reasonably used to measure the *object-ness* of a proposal. In recent years, Lee *et al.* [16] introduced an alternative clustering process, Ma and Latecki [19] attempted to model the selection process as a constrained maximum weight cliques problem, and Zhang *et al.* [32] proposed a layered directed acyclic graph based framework.

III. Our Approach

A. Overview

Our goal is to jointly segment multiple videos containing a common object in an unsupervised manner. We consider this task as an object optimization process consists of *object discovery*, *object refinement* and *object segmentation* executed on the whole set of videos. In this optimization process, we use a spatio-temporal SIFT flow that integrates optical flow, which captures inter-frame motion, and conventional SIFT flow, which captures across-videos correspondence information.

Our algorithm has three main stages: object discovery among multiple videos, object refinement between video pairs, and object segmentation on each video sequence.

Object Discovery: We use saliency and spatio-temporal SIFT flow to estimate common object regions in the entire video dataset. In this stage, an initial assignment of pixels belongs to object is performed.

Object Refinement: The goal is to refine the estimated object regions generated by prior step. This object refinement process is executed across a pairs of videos.

Object Segmentation: Since the correct estimation for object in each video is available, we can model the appearance of foreground and make segmentation on each video sequence to get more accurate results.

B. Object Discovery

In this stage, our method explores the video dataset structure and associates the global information with the intra-frame information like saliency to discover the common object from multiple videos, even in the presence of some frames without the common object. Three main properties of targeted object are helpful for object discovery: a) intra-frame saliency–the pixels of foreground should be relatively dissimilar to other pixels within a frame; b) inter-frame consistency–the pixels of foreground should be more consistent within a video; c) across-video similarity–the pixels of foreground should be more similar to other pixels between different videos (with possible changes in color, size and position). We propose a new spatio-temporal SIFT flow algorithm that integrates saliency, SIFT flow and optical flow to explore the correspondences between different videos. Thus, an object discovery energy function is then designed to effectively infer the common objects without the constraints that the object must exist in each frame. An overview of our algorithm is shown in Fig. 2.

Saliency of a pixel reflects how salient the pixel is, namely, the degree of its dissimilarity within the image. There are several methods in computer vision that concentrate on this topic. We use [24] yet any other saliency methods such as [23] can be incorporated. Let $\mathbf{V} = \{V_1, V_2, ..., V_N\}$ be a set of N input videos. $\mathbf{F}_n = \{F_n^1, F_n^2, ..., F_n^i, ...\}$ is a set of frames belong to video V_n. We compute a normalized

saliency map M_n^i for frame F_n^i. Based on intra-frame saliency property, the larger value of $M_n^i(\mathbf{x})$, the more likely that the pixel $\mathbf{x} = (x, y)$ belongs to object. Then we build a saliency term $\mathcal{A}_n^i(\mathbf{x})$ to define the cost of labeling pixel $\mathbf{x}$ for foreground ($l_n^i(\mathbf{x}) = 1$) or background ($l_n^i(\mathbf{x}) = 0$):

$$\mathcal{A}_n^i(\mathbf{x}) = exp - \{M_n^i(\mathbf{x})\} \cdot l_n^i(\mathbf{x}) + exp - \{1 - M_n^i(\mathbf{x})\} \cdot (1 - l_n^i(\mathbf{x})). \quad (1)$$

Optical flow [7] is represented as a 2D vector, which reflects the motion information of pixel $\mathbf{x}$ based on the color consistency assumption between consecutive frames. Optical flow algorithms can be used to estimate the inter-frame motion at each pixel in a video sequence. Let $\mathbf{v}_n^i$ denote the flow field between frame F_n^i and F_n^{i+1}. Here, a pixel $\mathbf{x}$ and its motion compensated pixel $\mathbf{x} + \mathbf{v}_n^i(\mathbf{x})$ are similar between two consecutive frames F_n^i and F_n^{i+1}, which represents the inter-frame consistency property. However, correspondences between object pixels in different videos could not be computed by optical flow since regions corresponding to the same object in different videos change in color, shape and position, which conflicts with the basic assumption of optical flow.

As an alternative, SIFT flow [17], [18] can be used to build a dense correspondence map across different scenes and object appearances. SIFT flow is shown to accommodate variations. We combine optical flow and local saliency into a superior spatio-temporal SIFT flow to build dense correspondences between pixels in different videos. Through spatio-temporal SIFT flow, reliable correspondences $\mathbf{w}_{nn'}^{ii'} = (u_{nn'}^{ii'}, v_{nn'}^{ii'})$ between the pixels of frame F_n^i and $F_{n'}^{i'}$ from different videos are established. In other words, pixel $\mathbf{x}$ of frame F_n^i is associated with the pixel $\mathbf{x}+\mathbf{w}_{nn'}^{ii'}(\mathbf{x})$ of frame $F_{n'}^{i'}$. These correspondences indicate whether pixels belong to the common object (even when it may be very salient within its frame).

We establish correspondences between a part of the pixels with high saliency values of one frame and the pixels from the frame of the other video. We select the pixels $\mathbf{R}_n^i = \{\mathbf{x} | M_n^i(\mathbf{x}) > \tau\}$ to explore their correspondences. In experiments, we fixed $\tau = 0.4$. This strategy improves matching accuracy by reducing the disturbance of those un-salient pixels which are very close to background, and it enables our method to remove some salient pixels that do not belong to common object.

Let s_n^i and $s_{n'}^{i'}$ be two SIFT fields of frame F_n^i and $F_{n'}^{i'}$ respectively that we want to match. The terms s_n^{i+1} and $s_{n'}^{i'+1}$ refer to the SIFT fields of frame F_n^{i+1} and $F_{n'}^{i'+1}$ respectively. F_n^{i+1} is the consecutive frame for F_n^i, and $\mathbf{N}_s$ is the spatial 8-neighborhoods of a pixel. Given the set of salient pixels $\mathbf{R}_n^i$, the energy function for spatio-temporal SIFT flow is defined as follows:

$$E = E_S + \alpha_1 E_{OS} + \alpha_2 E_{Disp} + \alpha_3 E_{Smooth} + \alpha_4 E_{Sal} \quad (2)$$

where the energy function contains the SIFT based data term

$$E_S(\mathbf{w}_{nn'}^{ii'}) = \sum_{\mathbf{x} \in \mathbf{R}_n^i} \left\| s_n^i(\mathbf{x}) - s_{n'}^{i'}(\mathbf{x}\mathbf{w}_{nn'}^{ii'}) \right\|_1, \quad (3)$$

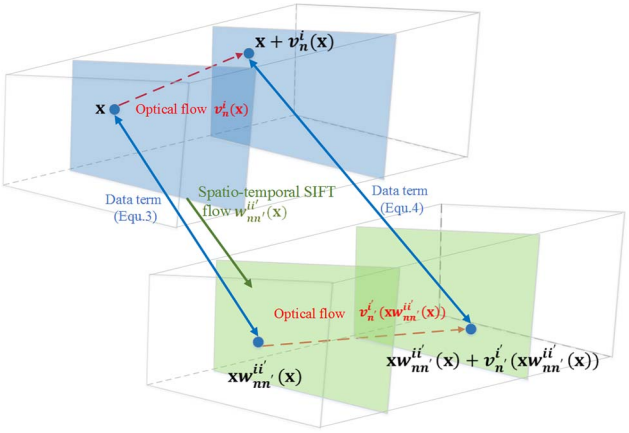


Fig. 3. Illustration of the data term ((3) and (4)) in the proposed spatio-temporal SIFT flow energy.

the optical flow compensated SIFT based data term

$$E_{OS}(\mathbf{w}_{nn'}^{ii'}) = \sum_{\mathbf{x} \in \mathbf{R}_n^i} \left\| s_n^{i+1}(\mathbf{x} + \mathbf{v}_n^i(\mathbf{x})) - s_{n'}^{i'+1}\left(\mathbf{x}\mathbf{w}_{nn'}^{ii'} + \mathbf{v}_{n'}^{i'}(\mathbf{x}\mathbf{w}_{nn'}^{ii'})\right) \right\|_1, \quad (4)$$

displacement term

$$E_{Disp}(\mathbf{w}_{nn'}^{ii'}) = \sum_{\mathbf{x} \in \mathbf{R}_n^i} \left\{ |u_{nn'}^{ii'}(\mathbf{x})| + |v_{nn'}^{ii'}(\mathbf{x})| \right\}, \quad (5)$$

the smoothness term

$$E_{Smooth}(\mathbf{w}_{nn'}^{ii'}) = \sum_{\substack{\mathbf{x},\mathbf{y} \in \mathbf{R}_n^i \\ \mathbf{x},\mathbf{y} \in \mathbf{N}_s}} \min\left\{ |u_{nn'}^{ii'}(\mathbf{x}) - u_{nn'}^{ii'}(\mathbf{y})|, d \right\} + \min\left\{ |v_{nn'}^{ii'}(\mathbf{x}) - v_{nn'}^{ii'}(\mathbf{y})|, d \right\}, \quad (6)$$

the saliency term

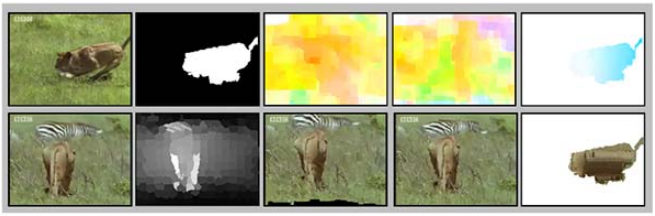
$$E_{Sal}(\mathbf{w}_{nn'}^{ii'}) = \sum_{\mathbf{x} \in \mathbf{R}_n^i} \left(1 - M_{n'}^{i'}(\mathbf{x}\mathbf{w}_{nn'}^{ii'}(\mathbf{x}))\right) + \left(1 - M_{n'}^{i'+1}(\mathbf{x}\mathbf{w}_{nn'}^{ii'}(\mathbf{x}) + \mathbf{v}_{n'}^{i'}(\mathbf{x}\mathbf{w}_{nn'}^{ii'}(\mathbf{x})))\right), \quad (7)$$

and we use the shorthand notation for SIFT matched pixels

$$\mathbf{x}\mathbf{w}_{nn'}^{ii'}(\mathbf{x}) = \mathbf{x} + \mathbf{w}_{nn'}^{ii'}(\mathbf{x}).$$

The data terms E_S and E_{OS} account for outliers in SIFT matching. The displacement term E_{Disp} models discontinuities of the pixel displacement field. The smoothness term E_{Smooth} employs one ℓ_1 norm to ensure the smoothness of field with the threshold d. This saliency constraint encourages matching the foreground pixels in frame F_n^i (F_n^{i+1}) with the pixels that have high saliency values in $F_{n'}^{i'}$ ($F_{n'}^{i'+1}$). Optical flow information is further introduced in data term (4). That is to say, if the SIFT descriptors of pixel $\mathbf{x}$ and $\mathbf{x}\mathbf{w}_{nn'}^{ii'}(\mathbf{x})$ have been matched by the data term in (3), the SIFT descriptors of pixels $\mathbf{x} + \mathbf{v}_n^i(\mathbf{x})$ and $\mathbf{x}\mathbf{w}_{nn'}^{ii'}(\mathbf{x}) + \mathbf{v}_{n'}^{i'}(\mathbf{x}\mathbf{w}_{nn'}^{ii'}(\mathbf{x}))$ on the optical flow direction should also be matched by the data term in (4).

Fig. 3 shows an indication for the data term ((3) and (4)) in the spatio-temporal SIFT flow function. Note that our



(a) (b) (c) (d) (e)

Fig. 4. Comparison between our spatio-temporal SIFT flow and traditional SIFT flow [17]. (a) A pair of frames need to be matched (top: frame F_n^i; bottom: frame $F_{n'}^{i'}$). (b) Saliency mask of frame F_n^i (the pixel $\mathbf{x}$ with $M_n^i(\mathbf{x}) > \tau$ is indicated as foreground) and saliency map M_n^i of frame $F_{n'}^{i'}$. (c) The SIFT flow and the result that frame $F_{n'}^{i'}$ warped onto frame F_n^i according to SIFT flow. The black region means the unsatisfying matching region outside the image range. (d) The spatio-temporal SIFT flow computed without saliency constraint, and the corresponding result that frame $F_{n'}^{i'}$ warped onto frame F_n^i. (e) The result of spatio-temporal SIFT flow by considering the saliency mask and saliency map in (b).

algorithm tries to match a portion of pixels (indicated by $\mathbf{R}_n^i$) instead of all the pixels within its frame in contrast to what the original SIFT flow [17] aims at. Belief propagation algorithms [17], [29] are applied to optimize above energy function.

Fig. 4 shows a comparison between the proposed spatio-temporal SIFT flow and traditional SIFT flow. Fig. 4(a) depicts two frames F_n^i and $F_{n'}^{i'}$ need to be matched. Fig. 4(b) are the computed saliency mask (of frame F_n^i) and the saliency map M_n^i of frame $F_{n'}^i$. Fig. 4(c) shows the result that frame $F_{n'}^{i'}$ warped onto frame F_n^i according to traditional SIFT flow. The black region is the matched area outside the image range, which is incorrect. Fig. 4(d) gives the result of spatio-temporal SIFT flow without saliency constraint. It is visible that spatio-temporal SIFT flow is more accurate than the conventional SIFT flow. Still, the performance of matching is not sufficient enough due to the disturbance of the background. The correct result should be that a 'lion' likes the one in frame F_n^i is presented in frame $F_{n'}^{i'}$. Fig. 4(e) shows the result of spatio-temporal SIFT flow by considering the saliency mask in Fig. 4(b), where the performance gains significant improvement.

Instead of using all the frames, it is possible to sample only a few representative frames or sample at a low frame-rate from video to perform the object discovery process (Fig. 2(a)). We select frame $\mathbf{f}_n = \{f_n^1, f_n^2, ..., f_n^k, ...\}$ every other five or ten frames from video V_n to perform object discovery process. For the *k-th* frame $f_1^k, f_2^k, ..., f_N^k$ of every video, we compute their spatio-temporal SIFT flow to capture their correspondence (Fig. 2(b)). Next, we calculate the distance of the point $\mathbf{x}$ of frame f_n^k from its corresponding points of other frames $\aleph(f_n^k) = \{f_1^k, ..., f_{n-1}^k, f_{n+1}^k, ...f_N^k\}$ in SIFT feature:

$$S_n^k(\mathbf{x}) = \frac{1}{|N-1|} \sum_{f_{n'}^k \in \aleph(f_n^k)} \|s_n^k(\mathbf{x}) - s_{n'}^k(\mathbf{x} + w_{nn'}^{kk}(\mathbf{x}))\|_1. \quad (8)$$

We normalize this term with values in [0, 1], where the smaller values indicate greater chance belonging to common object since the smaller distances to corresponding points.

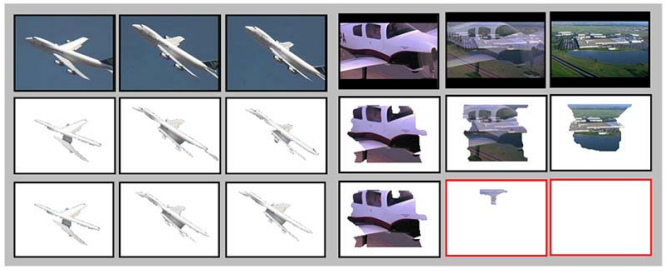


Fig. 5. Effective object discovery from multiple videos even with some frames not containing the common object. The first row shows two related video sequences and the common object **plane** does not appear in every frame. The object-like area of each frame estimated through (10) are presented in the second row. The bottom row shows the more correct object discovery results through (13) with further utilizing the inter-frame consistence property. Those frames with the ratio $\kappa \leq 0.2$ are considered not to contain the common object, which are marked in the red rectangles.

Similar to the saliency term, we build a matching term $\mathcal{M}_n^k(\mathbf{x})$ to define the cost of labeling pixel $\mathbf{x}$ for foreground ($l_n^k(\mathbf{x}) = 1$) or background ($l_n^k(\mathbf{x}) = 0$):

$$\begin{aligned}\mathcal{M}_n^k(\mathbf{x}) = & exp - \{S_n^k(\mathbf{x})\} \cdot l_n^k(\mathbf{x}) \\ & + exp - \{1 - S_n^k(\mathbf{x})\} \cdot (1 - l_n^k(\mathbf{x})). \end{aligned} \quad (9)$$

For frame f_n^k, we use the above saliency and matching terms to build an object discovery energy function as:

$$\mathcal{E}_n^k(\mathbf{x}) = \epsilon_1 \mathcal{A}_n^k(\mathbf{x}) + \epsilon_2 \mathcal{M}_n^k(\mathbf{x}) + \mathcal{V}_n^k(\mathbf{x}), \quad (10)$$

where the smooth term $\mathcal{V}_n^k(\mathbf{x})$ for frame f_n^k is expressed as:

$$\mathcal{V}_n^k(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x},\mathbf{y} \in \mathbf{N}_s} \|C_n^k(\mathbf{x}) - C_n^k(\mathbf{y})\|_2 \cdot |l_n^k(\mathbf{x}) - l_n^k(\mathbf{y})|, \quad (11)$$

where $C_n^k(\mathbf{x})$ indicates the color value of pixel $\mathbf{x}$ in f_n^k, spatial pixel neighborhood $\mathbf{N}_s$ consists of eight spatially neighboring pixels within one frame. This object discovery energy can be efficiently solved by traditional graph cut algorithm [27] and we are able to roughly estimate the common object over the video dataset. The scalars ϵ weight the various terms.

Effective object discovery from multiple videos even with some frames not containing the common object. The first row shows two related video sequences where the common object **plane** does not appear in every frame. The object-like area of each frame estimated through (10) is presented in the second row. The bottom row shows more accurate object discovery results through (13) with further utilizing the inter-frame consistence property. Those frames with the ratio $\kappa \leq 0.2$ are considered not to contain the common object, which are marked in the red rectangles.

There are many videos that include frames that do not contain the common object (e.g. the first row of Fig. 5). Current video co-segmentation approaches disregard this challenge and assume common object appears in every frame. Our method effectively handles this difficulty. One intuition is that the frames that do not contain the common object are not consistent with the frames that contain the object. Therefore, we further leverage the inter-frame consistency property. Based on (10), we get object-like areas and background areas for each frame. Suppose frame f_n^{k-1} contains the common foreground while f_n^k does not. Their estimated object-like

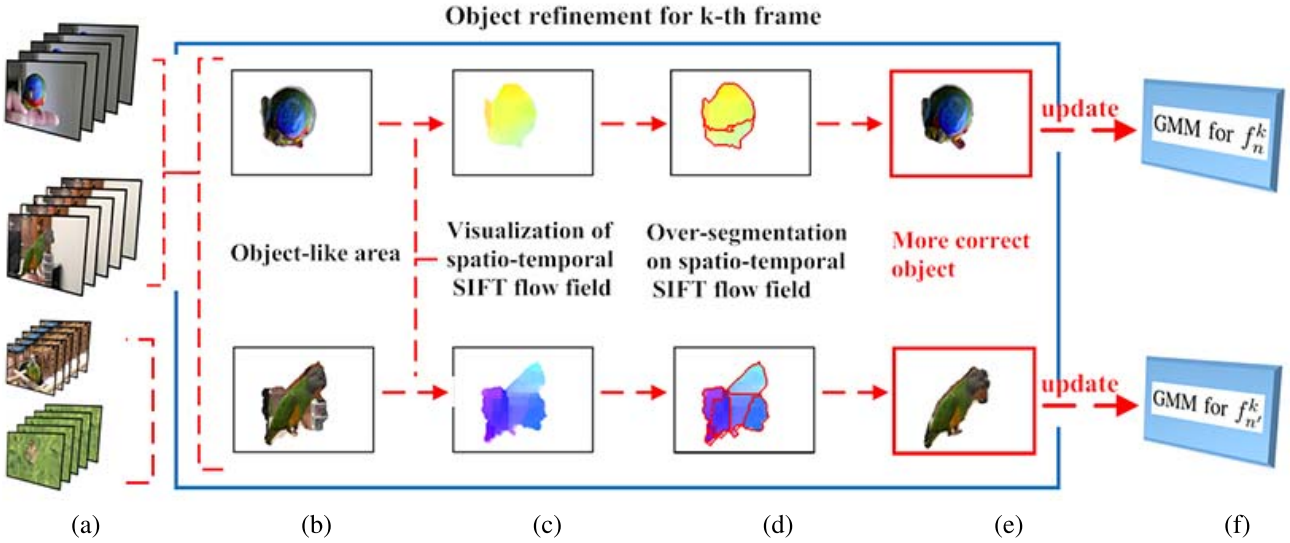


Fig. 6. Overview of our object refinement stage on frame f^k and frame f^{k+1}. (a) After object discovery step, a pair of videos is randomly selected to perform object refinement. (b) Object-like area is obtained after the object discovery step. (c) Visualization of spatio-temporal SIFT flow field. The discontinuities of spatio-temporal SIFT flow field reveal the variation of object structure. (d) Result of over-segmentation on spatio-temporal SIFT flow field. (e) A more accurate object partitioning is obtained by removing the pixels that are similar to background. (f) GMM for k^{th} frame is updated based on the updated estimation in (e).

area should be different. We employ Gaussian mixture models (GMM) to characterize the common object appearance. For frame f_n^{k-1}, the GMMs for object-like area and background are defined as $\{GMM^f_{f_n^{k-1}}, GMM^b_{f_n^{k-1}}\}$, respectively. We introduce an object consistence term to measure the consistency of estimated objects in video according to the appearance model of object. For frame f_n^k, this object consistence term is defined as:

$$\begin{aligned}\mathcal{C}_n^k(\mathbf{x}) = exp - \{p_n^k(\mathbf{x})\} \cdot l_n^k(\mathbf{x}) \\ + exp - \{1 - p_n^k(\mathbf{x})\} \cdot (1 - l_n^k(\mathbf{x})),\end{aligned} \tag{12}$$

where $p_n^k(\mathbf{x})$ denotes the probability of pixel $\mathbf{x}$ for foreground, which is obtained from $\{GMM^f_{f_n^{k-1}}, GMM^b_{f_n^{k-1}}\}$ of prior frame f_n^{k-1}.

Then we add this object consistence term into our object discovery energy function:

$$\mathcal{E}_n^k(\mathbf{x}) = \epsilon_1 \mathcal{A}_n^k(\mathbf{x}) + \epsilon_2 \mathcal{M}_n^k(\mathbf{x}) + \epsilon_3 \mathcal{C}_n^k(\mathbf{x}) + \mathcal{V}_n^k(\mathbf{x}). \tag{13}$$

We set parameter $\epsilon_1 = \epsilon_2 = \epsilon_3 = 50$ for all the test videos in our experiments. Since five or ten frames between frame f_n^{k-1} and f_n^k, the estimated GMM for frame f_n^{k-1} is helpful for identifying whether the frame f_n^k contains the common object. From Fig. 5 we see that the object discovery energy function in (13) is a better choice for detecting the frames not containing common object due to the inter-frame consistency.

We use $\mathbf{T}_n^k$ to denote the object-like area in frame f_n^k and the number of pixels belonging to the object-like area $\mathbf{T}_n^k$ is expressed as $|\mathbf{T}_n^k|$. We consider whether frame f_n^k contains the common object in case the ratio

$$\kappa_n^k = |\mathbf{T}_n^k| \,/\, |\mathbf{T}_n^{k-1}|, \tag{14}$$

is relatively large ($\kappa_n^k > 0.2$) and conclude that the foreground object of frame F_n^k is not changed. Conversely, if this ratio is small, we assume the objects between frame F_n^{k-1} and F_n^k are not consistent. In this case, frame F_n^k is considered to not contain the common object and we set $\mathbf{T}_n^k = \varnothing$. The GMM of the frame f_n^k is set as:

$$\begin{aligned}GMM^f_{f_n^k} &= GMM^f_{f_n^{k-1}}, \\ GMM^b_{f_n^k} &= GMM^b_{f_n^{k-1}}.\end{aligned}$$

In this way, the GMM for common object is kept consistent across the whole video sequence by ignoring the 'noise' frames. The frames that are detected to not contain the objects in object discovery step, will be not taken into consideration in next object refinement process.

C. Object Refinement

In the previous step, we obtain a coarse estimation for the common object in the dataset. Based on this, we seek to obtain a more accurate estimation for foreground object in every video. Our intuition is to remove the pixels that are similar to background based on the estimation result. Nevertheless, this also requires determining what foreground would look like. To filter out background pixels we divide the object-like area into sub-regions based on their variations. We utilize spatio-temporal SIFT flow for this purpose.

Fig. 6 illustrates the procedure of the object refinement step. First, a pair of videos $(V_n, V_{n'})$ is randomly selected from dataset. Their spatio-temporal SIFT flow between frames f_n^k and $f_{n'}^k$ is constructed. As shown in Fig. 6(c), discontinuities of spatio-temporal SIFT flow field reflect the variation of object structure (but not color variation) yet robust to object details. This property of spatio-temporal SIFT flow field is very important. Through the computation of the discontinuities of spatio-temporal SIFT flow field, we divide the object-like area into a few regions depending on the

structure variation. This enables us to estimate every part of the object-like area whether belongs to foreground using GMMs.

Properties of flow field boundaries reveal the physical cues of object as investigated in the past [20], [26]. In [20], an embedding discontinuity detector is proposed for localizing object boundaries in trajectory spectral embedding, however this is not suitable for our work. In [26], an algorithm is presented to detect the motion boundary and determine which pixels reside inside the moving object is presented. This method faces difficulty when the foreground motion patterns are not distinct. Moreover, it divides the frame only into two parts, while we want to divide the object-like area into multiple regions based on the structure variations.

Based on the visualization of spatio-temporal SIFT flow field using [1], numerous over-segmentation methods can be introduced and the object-like area can be efficiently partitioned into regions as shown in Fig. 6(d). Each pixel denotes a flow vector where the orientation and magnitude are represented by the hue and saturation of the pixel, respectively.

For each region $\mathbf{t}$ of object-like area $\mathbf{T}_n^k$, we build the $GMM_{\mathbf{t}}^b$ for background $\overline{\mathbf{T}}_n^k$ and the $GMM_{\mathbf{t}}^f$ for the remaining region (object) $\mathbf{T}_n^k \setminus \mathbf{t}$. The likelihood $\rho_n^k(\mathbf{x}_t)$ of pixels $\mathbf{x}_t \in \mathbf{t}$ for foreground is estimated using $\{GMM_{\mathbf{t}}^f, GMM_{\mathbf{t}}^b\}$.

We compare the texture of region $\mathbf{t}$ with the background and object-like area using the local binary pattern (LBP) features, which is used for describing the local spatial structure of an image. To model the texture of foreground and background in frame f_n^k, two normalized histograms ($H_{\mathbf{t}}^f$ and $H_{\mathbf{t}}^b$) are calculated in LBP domain. For region $\mathbf{t}$, the pixels belonging to the object-like area $\mathbf{T}_n^k \backslash \mathbf{t}$ are used for formulating the LBP histogram $H_{\mathbf{t}}^f$ while the pixels belonging to the background area $\overline{\mathbf{T}}_n^k$ are sampled for forming $H_{\mathbf{t}}^b$. Thus the probability $\ell_n^k(\mathbf{x}_t)$ of pixels $\mathbf{x}_t \in \mathbf{t}$ for foreground is estimated through these two LBP histograms as follows:

$$\ell_n^k(\mathbf{x}_t) = \frac{H_{\mathbf{t}}^f[\mathbf{x}_t]}{H_{\mathbf{t}}^f[\mathbf{x}_t] + H_{\mathbf{t}}^b[\mathbf{x}_t]}, \tag{15}$$

where $H_{\mathbf{t}}^f[\mathbf{x}_t]$ (with value in [0, 1]) indicates the value of histogram $H_{\mathbf{t}}^f$ at pixel $\mathbf{x}_t$.

We combine $\rho_n^k(\mathbf{x}_t)$ and $\ell_n^k(\mathbf{x}_t)$ as follows:

$$o_n^k(\mathbf{x}_t) = \beta \cdot \rho_n^k(\mathbf{x}_t) + (1-\beta) \cdot \ell_n^k(\mathbf{x}_t) \quad 0 < \beta < 1, \tag{16}$$

where the term $o_n^k(\mathbf{x}_t)$ denotes the probability of the pixel $\mathbf{x}_t$ for foreground according to both appearance and texture models. If $o_n^k(\mathbf{x}_t) < 0.5$, pixel $\mathbf{x}_t$ will be classified into background. There is no need to consider all of regions $\mathbf{t} \in \mathbf{T}_n^k$. If the area of region $\mathbf{t}$ is too large or too small, we will ignore these regions. These constraints will take fewer regions into account and enhance the efficiency of our object refinement. In our experiments, the region with $\frac{|\mathbf{t}|}{|\mathbf{T}_n^k|} > 0.5$ or $\frac{|\mathbf{t}|}{|\mathbf{T}_n^k|} < 0.05$ will be directly classified into foreground.

After frame f_n^k has been refined, we update $\{GMM_{f_n^k}^f, GMM_{f_n^k}^b\}$ (Fig. 6(f)) to provide guidance for the following object segmentation process. As shown in Fig. 6, this object refinement process is executed across video pairs and more correct estimation for foreground object is achieved.

D. Object Segmentation by Optimization

Once the correct estimations for foreground of each video are obtained, a graph-cut based method is employed to get per-pixel segmentation results. Recall our definition of $\mathbf{f}_n = \{f_n^1, f_n^2, ..., f_n^k, ...\}$ is that we select frame $\mathbf{f}_n$ every other five or ten frames from video V_n. After the object refinement process, we get more correct estimation for common object and update the appearance model of the object and background $\{GMM_{f_n^k}^f, GMM_{f_n^k}^b\}$ for frame f_n^k, which can be used to conduct the segmentation in next five or ten frames of f_n^k. For frame F_n^i, we obtain the likelihood of pixel $\mathbf{x}$ for foreground as $p_n^i(\mathbf{x})$ using our appearance models estimated by its temporally nearest frame of $\mathbf{f}_n$.

For video V_n, we update the labelling $\{l_n^i\}_i$ for all pixels to obtain the final segmentation results through an object segmentation function. This object segmentation function $\mathcal{F}_n(\mathbf{x})$ based on spatio-temporal graph by connecting frames temporally can be defined as follows:

$$\mathcal{F}_n(\mathbf{x}) = \sum_i \{\sum_{\mathbf{x}} \mathcal{U}_n^i(\mathbf{x}) + \gamma_1 \sum_{\mathbf{x},\mathbf{y}\in\mathbf{N}_s} \mathcal{V}_n^i(\mathbf{x},\mathbf{y}) + \gamma_2 \sum_{\mathbf{x},\mathbf{y}\in\mathbf{N}_t} \mathcal{W}_n^i(\mathbf{x},\mathbf{y})\}, \tag{17}$$

where the set $\mathbf{N}_s$ contains all the 8-neighbors within one frame and the set $\mathbf{N}_t$ contains the backward nine neighbors in pairs of adjacent frames. The parameters γ are the positive coefficient for balancing the relative influence between various terms.

The unary term $\mathcal{U}_n^i$ defines the cost of labeling pixel $\mathbf{x}$ with foreground and background according to our appearance model:

$$\begin{aligned}\mathcal{U}_n^i(\mathbf{x}) = {} & exp - \{p_n^i(\mathbf{x})\} \cdot l_n^i(\mathbf{x}) \\ & + exp - \{1 - p_n^i(\mathbf{x})\} \cdot (1 - l_n^i(\mathbf{x})).\end{aligned} \tag{18}$$

where $p_n^i(\mathbf{x})$ denotes the probability of pixel $\mathbf{x}$ for foreground as we mentioned before. The pairwise terms $\mathcal{V}_n^i$ and $\mathcal{W}_n^i$ encourage spatial and temporal smoothness, respectively. These two terms favor assigning the same label to neighboring pixels that have similar color:

$$\begin{aligned}\mathcal{V}_n^i(\mathbf{x},\mathbf{y}) &= \sum_{\mathbf{x},\mathbf{y}\in\mathbf{N}_s} \|C_n^i(\mathbf{x}) - C_n^i(\mathbf{y})\|_2 \cdot |l_n^i(\mathbf{x}) - l_n^i(\mathbf{y})|, \\ \mathcal{W}_n^i(\mathbf{x},\mathbf{y}) &= \sum_{\mathbf{x},\mathbf{y}\in\mathbf{N}_t} \|C_n^i(\mathbf{x}) - C_n^{i+1}(\mathbf{y})\|_2 \cdot |l_n^i(\mathbf{x}) - l_n^{i+1}(\mathbf{y})|.\end{aligned} \tag{19}$$

We use binary graph cuts [27] to obtain the optimal solution for (17), and thus get the final segmentation results. The final labelling $\{l_n^i\}_i$ for all pixels in all frames represents a segmentation of the video V_n.

IV. Experimental Results

A. ViCoSeg Dataset

The purpose of this work is to automatically co-segment the common objects from related videos with large foreground/background motion patterns or appearance variations, even when some frames do not contain the common object. There has been very little comparative work to address these problems. To deeper explore these issues and establish

TABLE I
DATASET STATISTICS

Method	Groups	Videos	Foreground(s)
Rubio *et al.* [21]	4	13	Single
Chiu and Fritz [25]	4	11	Multiple
our ViCoSeg	12	30	Single

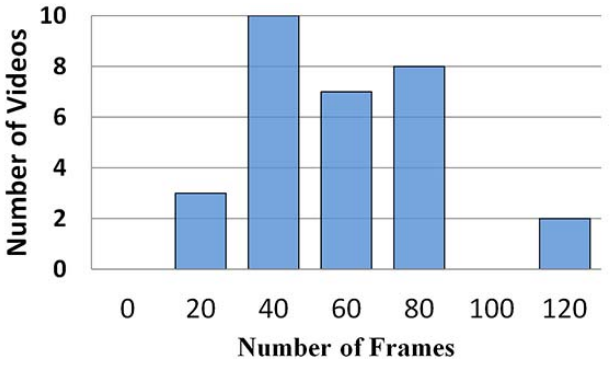


Fig. 7. The histogram of the number of frames in videos.

a benchmark for future work, we introduce a video co-segmentation dataset, called *ViCoSeg*, which is collected from existing databases and Youtube with similar characteristics in order to perform co-segmentation. While previous works have experimented with a few of videos, this dataset consists of 12 groups of videos including 30 videos totally, together with their corresponding pixel-level ground truth. Each group of videos includes two to four video clips. Table 1 lists some statistics of our introduced dataset. We note that this dataset is significantly larger than those used in previous works (Rubio *et al.* [21] and Chiu and Fritz [25]). Chiu and Fritz [25] proposed a video co-segmentation dataset with multiple objects, which is not very suitable for the task of single object video co-segmentation. This dataset is also limited with the number of video groups. Rubio *et al.* [21] offered a dataset for single object video co-segmentation. This dataset is not satisfactory since it only consists of 4 video groups, and the same foreground is simply pasted into different backgrounds.

Our proposed dataset is the largest co-segmentation dataset as far as we know, and the number of frames in each video is also shown in Fig. 7. These selected video sequences range in length from 20 to 125 frames and exhibit major challenges such as foreground/background color overlap (e.g. **Tiger**), large shape deformation (e.g. **Gokart**), and various motion patterns (e.g. **Boat**), etc. Four groups (**Gokart**, **Lion**, **Horse**, **Tiger**) of our dataset have similar objects. Different from previous datasets, the introduced video groups include six video groups (**Bird**, **Boat**, **Car**, **Cat**, **Moto**, **Plane**) that have large intra-class variations. In fact, most previous works assume the common objects share similar appearance model. Moreover, the **Car2** and **Plane2** contain some frames without the common object, which increases difficulties for the co-segmentation task. In particular, there are very few methods emphasis on this issue. We will further evaluate the performance of our proposed approach in these scenes and give detailed discussion in the following subsections.

B. Experiments

We have tested our method on four video groups with similar foreground from our ViCoSeg dataset. In order to demonstrate the effectiveness of our algorithm on segmenting out the common object from diverse categories, we further test on video groups with large variations on foreground class. We finally evaluate our method on **Car2** and **Plane2** video groups, these two video groups have some frames not containing the common object.

We present qualitative video co-segmentation results and comparisons with multiple-class video co-segmentation [25] and video object segmentation [32]. [25] proposed a multi-class video co-segmentation method using a non-parametric Bayesian model. Since the video dataset is divided into multi-class, we select the class that has the most overlap with the ground truth as its foreground segmentation result. [32] proposed an approach to extract primary object segments in videos in the object proposal domain by combining motion, appearance and predicted-shape similarity across frames. We use the publicly available implementations by authors of these methods and set their free parameters so as to maximize their performance for fairness.

We also present quantitative comparisons with previous methods [25], [32]. In our experiments, two main metrics are employed for the evaluation. Segmentation performance is measured by the *average per-frame pixel error* [11], which is the number of falsely labeled pixels both foreground and background. This measurement is defined as $\frac{|\mathbf{XOR}(R,GT)|}{K}$, where R indicates the segmentation result, GT and K correspond to ground truth and the number of frames respectively. Additionally, we adopt the *intersection-over-union score* [25] for evaluation, which is the standard in PASCAL challenges and defined as $\frac{R \cap GT}{R \cup GT}$ as criterion. Since both accuracy and temporal coherence are very important for video segmentation, we further provide the demo video of co-segmentation in our website[2] for demonstrating the temporal coherence of the proposed approach.

C. Video Co-Segmentation With Similar Foreground

We collect four video groups that have same or similar object, and also a pixel-level segmentation ground-truth for each video is available. Our approach is evaluated on these video groups and compared to object-based video segmentation (OS) [32] and multi-class video co-segmentation (MC) [25]. Fig. 8 shows the comparison results between our algorithm and previous methods. Each group of the categories **Gokart** and **Lion** consists of two source videos, while each of the rest ones comprises three videos.

OS [32] extracts primary object for single video segmentation using object proposals. Their method presented a motion scoring function by optical flow gradients and shape similarity for selection of object proposals. However, the performance becomes not well in complex videos such as **Tiger** video set. The visual similarity between background and foreground

[2]http://github.com/shenjianbing/robustvideocoseg/demo.mp4

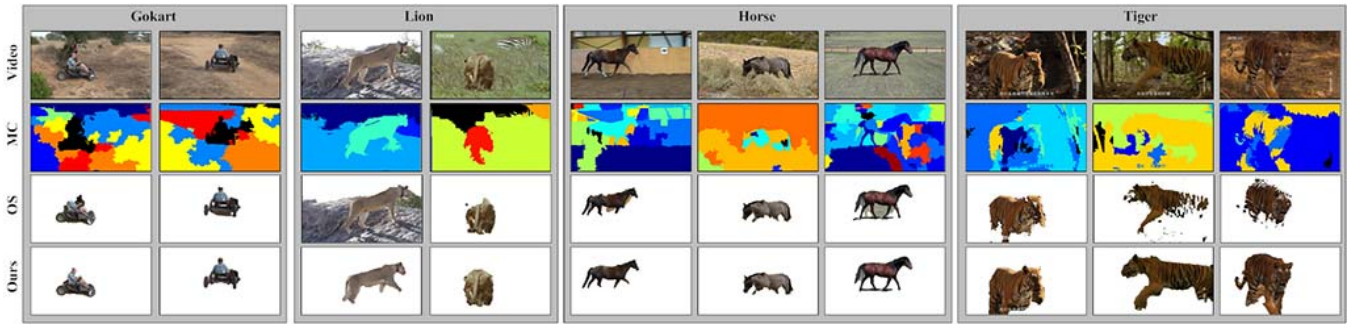


Fig. 8. Video co-segmentation results on four groups of videos with similar foreground. In each group, the segmentation results from top to bottom are generated by MC [25], OS [32] and our method, respectively.

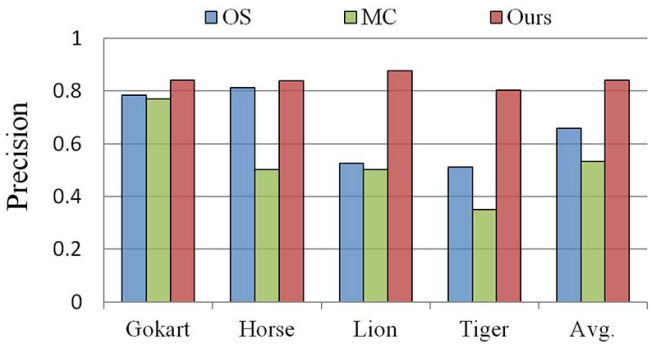


Fig. 9. The intersection-over-union metric on our video sets with similar foreground.

misleads the inference of common object, which causes the poor segmentation performance. The effective utilization of saliency, motion and SIFT features in our spatio-temporal field proves more correct indication for foreground object and leads to significant improvements.

MC [25] provided a way that combines global appearance and motion cues to perform multi-class video co-segmentation. However, their estimation for global appearance model relies on chroma and motion features. The discrimination power of this model is limited to objects with complicated appearance. For instance, in the videos of **Horse** and **Tiger**, their classification does not correspond to a particular object but only to regions that exhibit coherent appearance or motion.

Our method utilizes a deeper understanding of the properties of foreground object, including intra-frame saliency, inter-frame consistency and across-video similarity. These properties are further integrated into our spatio-temporal SIFT flow and object discovery energy function, which enables our method to produce more accurate results and outperform the state-of-the-art approaches. For **Lion** group, our method extracts the entire foreground lion from the backgrounds, although the zebra also looks like foreground. For **Tiger** group, the foreground tiger is not very distinctive from the background in terms of color, but it is still successfully discovered by our algorithm due to good correspondences to other related videos. Fig. 9 and Table. 2 report the quantitative comparisons by the intersection-over-union metric and the average per-frame pixel errors respectively. Experiment shows that our method produces much higher segmentation accuracy.

TABLE II
AVERAGE PER-FRAME PIXEL ERRORS

Method	OS[32]	MC[25]	Ours
Gokart	1126	3356	**701**
Horse	1781	13970	**1373**
Lion	21575	7584	**1351**
Tiger	11886	48607	**4857**
Avg.	9092	18429	**2071**

D. Video Co-Segmentation With Large Intra-Class Variations

Our framework can automatically produce object co-segmentation results for videos with unrelated backgrounds and is robust to the intra-class variation. To further illustrate this advantage of our algorithm, we collect 6 video groups with large variations in foreground appearance and apply our approach to co-segment these challenging video groups. As shown in Fig. 10, we compare our video co-segmentation results with the ones generated by previous methods OS [32] and MC [25]. Video groups of both **Bird** and **Boat** have four source videos, while others include a pair of two videos. Most foreground objects have little appearance similarities, which are very challenging for segmentation. Even so, the experimental results clearly show that the performance of our method is much better than OS [32] and MC [25] in this issue.

The segmented result by OS [32] is not accurate in many videos in Fig. 10. For example, in **Bird** videos, a lot part of background is wrongly divided into foreground bird together. That is because it lacks of considering inter-video object correspondence which is much helpful to predict meaningful segmentations. In contrast, our method infers the object of interest across videos based on spatio-temporal SIFT flow, which robustly discovers object over the entire database. Moreover, OS [32] emphasizes that optical flow gradients are able to discriminate objects and background. But in many scenes, object does not have distinct motions from the background. For example, in the group of **Boat** videos, the boat is moving and the river is flowing, the similar motion cues of object and background lead to incorrect segmentation results of OS [32].

As a video object co-segmentation method, MC [25] takes into account the correspondences of objects across videos. But this relationship is based on the similarity of

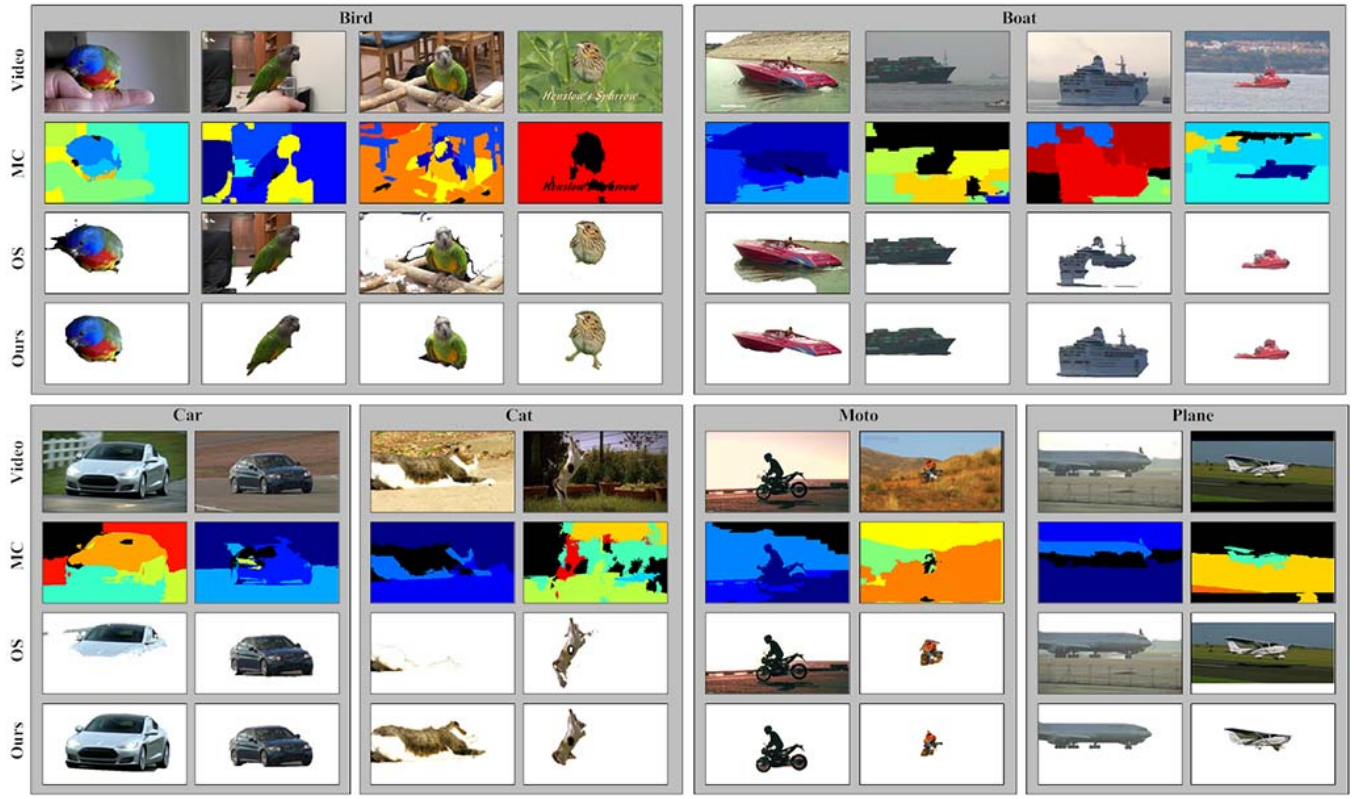


Fig. 10. Video co-segmentation results on six groups of videos with high intra-class variations. In each group, the results from top to bottom are generated by MC [25], OS [32] and our method, respectively.

object appearance and lacks of high level features. As a result, it does not correctly discriminate foreground across videos when the common object has appearance variations and wrongly merges object classes from the foreground and background. In **Cat** videos we have observed that a part of cat (black label) in the first video is wrongly classified together with the background tree in the second video. In addition, the lackness of high level features for common object brings difficulties in capturing the foreground object in its entirety. MC [25] produces fragment in class labeling, as shown in the third video of **Bird** and the second video of **Cat** in Fig. 10.

Our algorithm builds the correspondences of objects, incorporates object structure, texture related cues to capture the variability of object and refines the estimation results for object. In this way, our algorithm provides significant improvement over previous methods OS [32] and MC [25]. For instance, in **Car** videos, the foreground car in the first clip is obviously different from the one in the second clip, while our method outputs nearly perfect segmentations according to the reliable foreground correspondences by our spatio-temporal SIFT flow. For **Moto** group, there is large foreground scale change and different foreground appearance. For **Boat** group, the boat in the first clip moves fast while the one in the third video is almost static. Our method is robust enough to produce more satisfactory and accurate co-segmentation results, since our algorithm associates various cues for foreground object. The intersection-over-union metric is shown in Fig. 11 and the average per-frame pixel errors are illustrated in Table 3. With an overall average performance of 85.21% of our method in intersection-over-union metric measurement, we make significant improvement for co-segmentation results over OS (52.71%) and MC (44.74%).

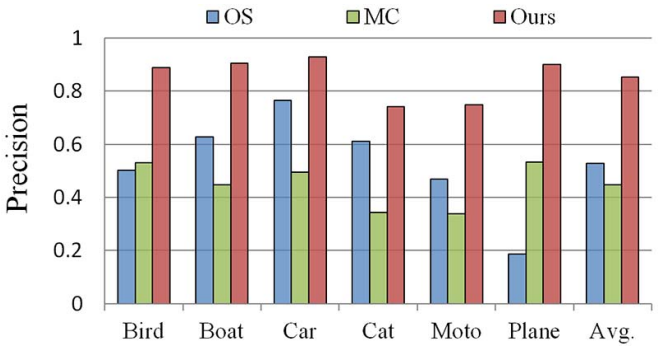


Fig. 11. The intersection-over-union metric on our video sets with large intra-class variations.

TABLE III
AVERAGE PER-FRAME PIXEL ERRORS

Method	OS[32]	MC[25]	Ours
Bird	12192	10483	**1178**
Boat	8240	9754	**972**
Car	5753	16192	**1638**
Cat	6636	18404	**2781**
Moto	13043	12672	**849**
Plane	47526	22808	**675**
Avg.	15565	15052	**1349**

E. Co-Segmenting Some Frames Without Common Object

It is very common that some frames do not contain the object of interest in real video data, such as object moving

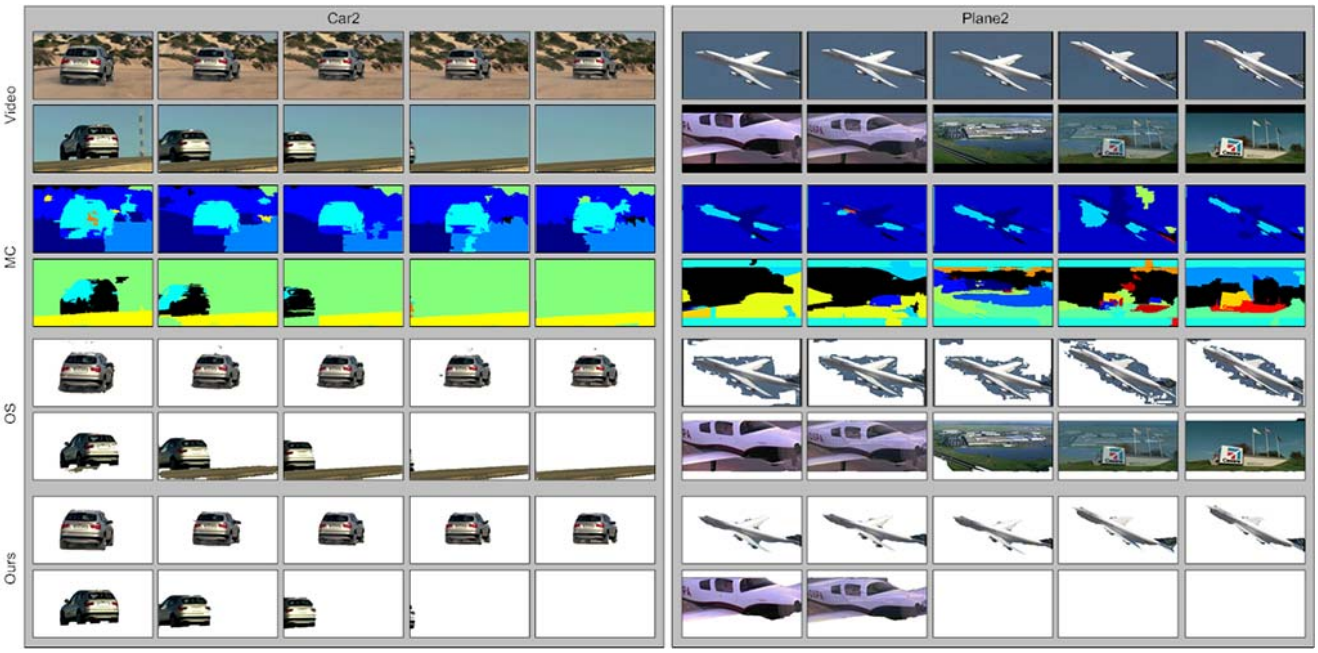


Fig. 12. Video co-segmentation results on two groups of videos with some frames not containing the common object. In each group, the results from top to bottom are generated by MC [25], OS [32] and our method, respectively.

out of camera or the shot switching effect. But there are very few methods notice this fact, most of co-segmentation methods assume every frame contains the interesting object, which cannot handle these issues well. The proposed method tries to tackle these problems and is evaluated on the newly collected two video groups: **Car2** and **Plane2**, since there are no suitable video co-segmentation datasets for this issue. Both **Car2** and **Plane2** have two source videos and some frames do not contain the foreground object. The frames without common object are naturally indicated by returning an empty labeling through our method. As shown in Fig. 12, our co-segmentation results clearly show that our method handles such situation better than OS [32] and MC [25], which maintains the good segmentation performance even though the objects are not appear in every frame.

The co-segmentation results by OS [32] are not satisfying, especially for the frames not containing object. For instance, background road is wrongly classified into foreground in **Car2** videos. The OS approach does not consider the situation that objects do not appear in each frame in the video clip. Moreover, we can find that OS does not work well for the large foreground object or with low contrast to background. This issue is particularly prominent in the second video in **Plane2**, where the plane occupies almost all the picture and is not distinct with background. That maybe because OS heavily relies on the initial proposals estimated for foreground. When the foreground is difficult to identify, the objectness measure for proposals is not accurate and segmentation errors then occur. MC [25] builds a global appearance model by considering the color information of different classes. Based on this effort, the segments of the same class are linked within and across videos. For **Car2**, all the backgrounds are correctly labeled for the frames not containing the foreground car. However, when the backgrounds are visually similar to object, the global appearance model will be limited for its weak ability of object discovery in the videos of **Plane2**.

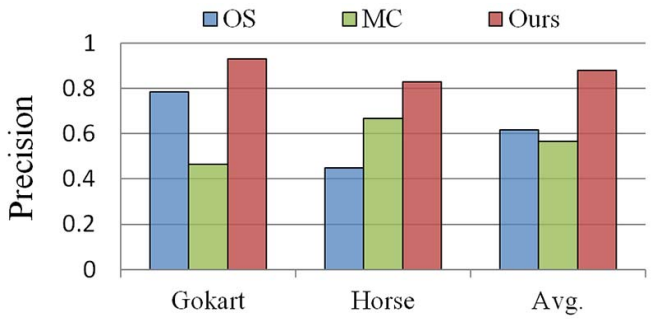


Fig. 13. The intersection-over-union metric on our video sets with large intra-class variations.

TABLE IV
AVERAGE PER-FRAME PIXEL ERRORS

Method	OS[32]	MC[25]	Ours
Car2	5715	5088	**670**
Plane2	34048	10095	**1798**
Avg.	19882	15183	**1234**

Different from previous co-segmentation methods, our algorithm emphasizes on the object discovery by incorporating more discriminative visual cues like SIFT feature and deeply exploring the correspondences between foreground objects within and across videos. Based on this effective inference for foreground, we build the consistent foreground appearance models across the whole video sequence. This strategy makes our method powerful enough for detecting the frames without object. As shown in Fig. 13, our method obviously obtains the better co-segmentation results with

more spatio-temporally consistency than the results by OS [32] and MC [25]. The intersection-over-union metric is shown in Fig. 13 and the average per-frame pixel errors are illustrated in Table 4. Both the quantitative and qualitative experimental results demonstrate that our method achieves much better co-segmentation results than the state-of-the-art OS [32] and MC [25] approaches.

V. Conclusion

We presented a robust video co-segmentation method that discovers the common object over an entire video dataset and segments out the objects from the complex backgrounds. Saliency, motion cues and SIFT flow are integrated into our spatio-temporal SIFT flow to explore the relationships between foreground objects. Furthermore, we formulate the video co-segmentation problem as an object optimization process, which progressively refine the estimation for object in three steps: object discovery, object refinement and object segmentation. Both the quantitative and qualitative experimental results have shown that the proposed algorithm creates more reliable and accurate video co-segmentation performance than the state-of-the-art algorithms. Unlike previous work, we emphasize that object discovery process should be robust to foreground variations in appearance or motion patterns, which extends the applicability of our co-segmentation method.

References

[1] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, Mar. 2011.

[2] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video SnapCut: Robust video object cutout using localized classifiers," *ACM Trans. Graph.*, vol. 28, no. 3, Aug. 2009, Art. ID 70.

[3] L. S. Silva and J. Scharcanski, "Video segmentation based on motion coherence of particles in a video sequence," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 1036–1049, Apr. 2010.

[4] Y. Huang, Q. Liu, and D. Metaxas, "Video object segmentation by hypergraph cut," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1738–1745.

[5] J. Yuen, B. Russell, C. Liu, and A. Torralba, "LabelMe video: Building a video database with human annotations," in *Proc. 12th IEEE ICCV*, Sep./Oct. 2009, pp. 1451–1458.

[6] W. Brendel and S. Todorovic, "Video object segmentation by tracking regions," in *Proc. 12th IEEE ICCV*, Sep./Oct. 2009, pp. 833–840.

[7] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[8] I. Endres and D. Hoiem, "Category independent object proposals," in *Proc. ECCV*, 2010, pp. 575–588.

[9] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. IEEE CVPR*, Jun. 2010, pp. 73–80.

[10] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 3241–3248.

[11] D. Tsai, M. Flagg, and J. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *Proc. BMVC*, 2010, pp. 56.1–56.11.

[12] T. Wang and J. Collomosse, "Probabilistic motion diffusion of labeling priors for coherent video segmentation," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 389–400, Apr. 2012.

[13] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 2141–2148.

[14] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller, "Multiple hypothesis video segmentation from superpixel flows," in *Proc. ECCV*, 2010, pp. 268–281.

[15] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. ECCV*, 2010, pp. 282–295.

[16] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proc. IEEE ICCV*, Nov. 2011, pp. 1995–2002.

[17] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[18] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in Internet images," in *Proc. IEEE CVPR*, Jun. 2013, pp. 1939–1946.

[19] T. Ma and L. J. Latecki, "Maximum weight cliques with mutex constraints for video object segmentation," in *Proc. IEEE CVPR*, Jun. 2012, pp. 670–677.

[20] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proc. IEEE CVPR*, Jun. 2012, pp. 1846–1853.

[21] J. C. Rubio, J. Serrat, and A. López, "Video Co-segmentation," in *Proc. ACCV*, 2012, pp. 13–24.

[22] D.-J. Chen, H.-T. Chen, and L.-W. Chang, "Video object cosegmentation," in *Proc. ACM Multimedia*, 2012, pp. 805–808.

[23] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[24] B. Jiang, L. Zhang, H. Lu, M.-H. Yang, and C. Yang, "Saliency detection via absorbing Markov chain," in *Proc. IEEE ICCV*, Dec. 2013, pp. 1665–1672.

[25] W.-C. Chiu and M. Fritz, "Multi-class video Co-segmentation with a generative multi-video model," in *Proc. IEEE CVPR*, Jun. 2013, pp. 321–328.

[26] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proc. IEEE ICCV*, Dec. 2013, pp. 1777–1784.

[27] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[28] Y. Li, J. Sun, and H.-Y. Shum, "Video object cut and paste," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 595–600, Jul. 2005.

[29] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, Oct. 2006.

[30] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. ECCV*, 2010, pp. 366–379.

[31] C. Xu, C. Xiong, and J. J. Corso, "Streaming hierarchical video segmentation," in *Proc. ECCV*, 2012, pp. 626–639.

[32] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *Proc. IEEE CVPR*, Jun. 2013, pp. 628–635.

[33] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE CVPR*, Jun. 2015, pp. 3395–3402.

Wenguan Wang is currently pursuing the Ph.D. degree with the School of Computer Science, Beijing Institute of Technology, Beijing, China. His current research interests include video segmentation and co-segmentation.

Jianbing Shen (M'11–SM'12) is currently a Professor with the School of Computer Science, Beijing Institute of Technology, Beijing, China. His research interests include computer vision and multimedia processing. He is on the Editorial Boards of the *Neurocomputing* journal.

Xuelong Li (M'02–SM'07–F'12) is currently a Professor with the Center for OPTical IMagery Analysis and Learning, School of Computer Science, Northwestern Polytechnical University, Xi'an, China.

Fatih Porikli (F'13) received the Ph.D. degree from New York University, New York, NY, USA, in 2002. He served as a Distinguished Research Scientist with Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. He has contributed broadly to object detection, motion estimation, tracking, image-based representations, and video analytics. He is currently a Professor with the Research School of Engineering, Australian National University, Canberra, ACT, Australia. He is also acting as the Leader of the Computer Vision Group with NICTA, Sydney, NSW, Australia.