Semantic Co-segmentation in Videos

Yi-Hsuan Tsai[1(✉)], Guangyu Zhong[1,2], and Ming-Hsuan Yang[1]

[1] UC Merced, Merced, USA
{ytsai2,gzhong,mhyang}@ucmerced.edu
[2] Dalian University of Technology, Dalian, China

Abstract. Discovering and segmenting objects in videos is a challenging task due to large variations of objects in appearances, deformed shapes and cluttered backgrounds. In this paper, we propose to segment objects and understand their visual semantics from a collection of videos that link to each other, which we refer to as *semantic co-segmentation*. Without any prior knowledge on videos, we first extract semantic objects and utilize a tracking-based approach to generate multiple object-like tracklets across the video. Each tracklet maintains temporally connected segments and is associated with a predicted category. To exploit rich information from other videos, we collect tracklets that are assigned to the same category from all videos, and co-select tracklets that belong to true objects by solving a submodular function. This function accounts for object properties such as appearances, shapes and motions, and hence facilitates the co-segmentation process. Experiments on three video object segmentation datasets show that the proposed algorithm performs favorably against the other state-of-the-art methods.

1 Introduction

Objects may appear at any location in various shapes and appearances with different visual semantics across videos. Given a set of videos, localizing and segmenting all the objects is a challenging task, especially when the visual categories are unknown. In this work, we propose an algorithm to segment objects and understand visual semantics from a video collection, which we refer to as *semantic co-segmentation*. Within the proposed co-segmentation framework, we aim to find the common representation for each semantic category and exploit relations between objects. For instance, dogs from different videos may share more commonalities and have stronger relations between each other than objects with other semantics (see Fig. 1).

Numerous algorithms have been proposed for video object co-segmentation [3,6,26,34]. However, most existing methods [3,6,26] assume that at least one common object appears all the time in two or more videos, which limits the

Y.-H. Tsai and G. Zhong—These authors contribute equally to this work.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46493-0_46) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part IV, LNCS 9908, pp. 760–775, 2016.
DOI: 10.1007/978-3-319-46493-0_46

applicability in real world scenarios. In this work, we propose an algorithm to segment semantic objects from a collection of videos containing various categories despite large variations in appearances, shapes, poses and sizes.

We exploit semantic information to facilitate co-segmentation to associate objects of the same category from different videos. Visual semantics has been used as prior information for object segmentation in weakly labeled videos [28,31,35]. In semantic video object segmentation, an object detector or a segmentation algorithm is first applied to localize objects according to the video label. However, for videos without any semantic label, an object detector may find noisy segments that do not belong to any semantic object (i.e., due to the trade-off between recall and precision). In this work, we propose an algorithm to associate semantic representations between objects in different videos and help the object co-segmentation process, where non-object detections can be removed.

Toward this end, we first extract semantic objects in each video. Compared with methods that use region proposals [34,35] to localize objects, we develop a proposal-free tracking-based approach that generates multiple tracklets of regions (segments) across the video. Each tracklet maintains temporal connections and contains a predicted category that is initialized by an image-based semantic segmentation algorithm. After collecting tracklets from all videos, we link the relations between tracklets for each object category by formulating a submodular optimization problem, which maximizes the similarities between object regions (segments). With this formulation, prominent objects in each video can be discovered and segmented based on similarities of regions.

We first conduct experiments on the Youtube-Objects dataset [22] in a weakly-supervised manner. Then we evaluate the proposed method in a more generalized setting without knowing any semantic information as a prior. Both results show that our algorithm performs favorably against the state-of-the-art methods. In addition, we compare our method to the other video object co-segmentation approaches on the MOViCS [3] and Safari [34] datasets. Experimental results on three datasets show that the proposed algorithm performs favorably in terms of visual quality and accuracy.

The contributions of this work are summarized as follows. First, we propose a semantic co-segmentation method that considers relations between objects from a collection of videos, where object categories can be unknown. Second, a proposal-free tracking-based method is developed to segment object-like tracklets while maintaining temporal consistency in videos. Third, a submodular function is formulated to carry out semantic co-segmentation from tracklets in all videos.

2 Related Work

Video Object Segmentation/Co-segmentation. Object segmentation from one single video has been studied extensively in the literature [10,14,15,21,29,33]. In general, these approaches are developed to use spatial-temporal graphical models based on object proposals [14,33], segments [15], motion cues [21] or propagating foreground regions [10,19,29]. Recently, co-segmentation methods

are developed to segment common objects in images [11,25,30] and videos [3,6,8,26,34]. Most co-segmentation schemes assume that all the input videos contain at least one common target object [3,6,8,26], which is rarely true in real world scenarios. With a less strict assumption in [34], objects with unknown number of categories can be segmented from a collection of videos by tracking and matching object proposals. However, another assumption underlying the above-mentioned methods is that usually common objects have almost identical appearances. In contrast, the proposed algorithm is not constrained by these factors and is able to segment objects with large variations in appearances without any assumption, e.g., number of object instances and number of object categories.

Object Segmentation in Weakly-Supervised Videos. Weakly-supervised methods have attracted attention due to their effectiveness for facilitating the segmentation process with known video-level object categories. Several learning-based approaches are proposed to collect semantic samples for training segment classifiers [9,28] or performing label transfer [16], and then identify the target object in videos. However, these methods rely on training instances and may generate inaccurate segmentation results. Zhang et al. [35] propose to segment semantic objects via detection without the need of training process. In this method, object detections and proposals are integrated within an optimization framework to refine the final tracklets for segmentation. In contrast, the proposed algorithm does not require object proposals or video-level annotations. More importantly, we link objects between different videos and construct a graph for submodular optimization, and hence help recognize each semantic object.

Object Discovery and Co-localization. Object discovery and co-localization methods are developed in a way similar to object co-segmentation, and these methods assume that input images or videos contain object instances from the same category. Recent image-based approaches [2,4,24,27] are proposed to overcome the problem of large amounts of intra-class variations and inter-class diversity. Several video-based methods are extended to account for temporal information. In [31], superpixel-level labels are propagated across frames via a boosting algorithm. However, this approach requires supervision from a few frame-level annotations. Kwak et al. [12] propose a video object discovery method by matching correspondences across videos and tracking object regions across frames. Different from the above-mentioned schemes, this work focuses on video object co-segmentation without any assumption on objects appearing in videos, in a way that we incorporate semantic information and analyze relations between object-like tracklets.

3 Proposed Algorithm

3.1 Overview

Given a set of videos with unknown object categories, our goal is to discover and segment prominent objects, as well as assign each object a semantic label. To achieve this, we first utilize a fully convolutional network (FCN) [17] trained

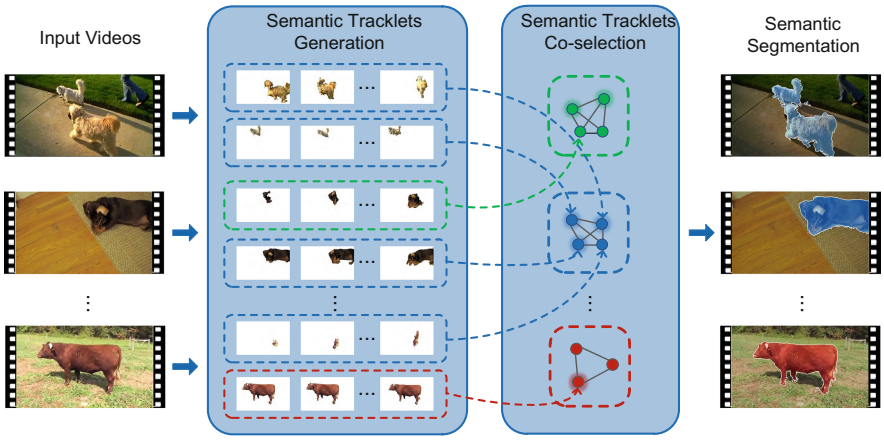

Fig. 1. Overview of the proposed algorithm. Given a collection of videos without providing category labels, we aim to segment semantic objects. First, a set of tracklets is generated for each video, and each tracklet is associated with a predicted category illustrated in different colors (e.g., blue represents the dog and red represents the cow). Then a graph that connects tracklets as nodes from all videos is constructed for each object category. We formulate it as the submodular optimization problem to co-select tracklets that belong to true objects (depicted as glowing nodes), and produce final semantic segmentation results. (Color figure online)

on the PASCAL VOC 2011 dataset [5] to segment objects in each frame, where each segment has a predicted category. To reduce noisy segments in each video, we cluster segments and eliminate clusters containing noisy segments through the video. Among the selected clusters with object segments, we randomly choose a few of them as initializations and apply a spatial-temporal graph-based tracking algorithm to generate tracklets. Each tracklet maintains coherent appearances of an object region (segment) in the spatial and temporal domains.

However, tracklets may still contain only object parts or noisy background clutters, and the available visual information is limited within each video. We construct a graph that connects tracklets within the same category from all videos as nodes, and utilize a submodular function to define the corresponding relations based on their appearances, shapes and motions. After maximizing this submodular function, tracklets are ranked according to their mutual similarities, and hence prominent objects can be discovered in each video. Figure 1 shows the overview of the proposed algorithm.

3.2 Semantic Tracklet Generation

Video object segmentation methods usually utilize object proposals in each frame to detect where instances may appear [6,14,15,34]. One challenge is to associate thousands of proposals from different objects while maintaining temporal connections for each of them across all sequences. Here, we propose to utilize a semantic segmentation algorithm (e.g., FCN) to generate object segments as

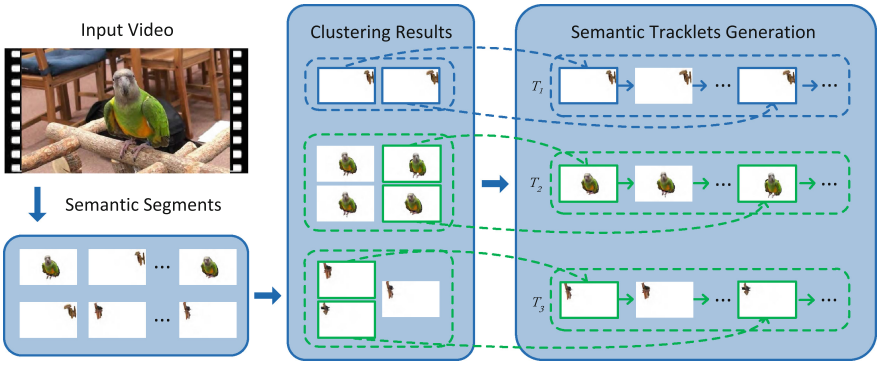

Fig. 2. Illustration of the proposed method for semantic tracklet generation. Given an input video, we first utilize the FCN algorithm to produce semantic segments in each frame. We then cluster all segments within each object category into different groups, where each color denotes one category (e.g., two green groups for birds and one blue group for dogs). Within each group, we randomly select a few segments as multiple initializations (depicted as rectangular boxes with solid color lines) and utilize a tracking-based approach to generate semantic tracklets T_i. Note that we only show the forward tracklets in this figure (similar process when generating backward tracklets). (Color figure online)

initializations, and then construct a spatial-temporal graphical model to track each object segment and form tracklets. The procedure to generate tracklets is illustrated in Fig. 2.

Selecting Objects Segments via Clustering. We first apply the FCN algorithm to extract object segments in each frame of one video. To reduce noisy segments that are not likely to be any object, a simple yet effective clustering method is utilized to select object-like segments through each video. Since the number of object instances is unknown, we apply the mean shift clustering method on all the segments within each object category based on color histograms in the RGB space. Then we select the N largest clusters (i.e., top 80 % of the largest ones) while removing the others.

The object segments in selected clusters are considered as initializations for tracking. We randomly choose a few segments from each cluster, while ensuring the selected segments are within a certain time frame (e.g., at least 20 frames apart between two selected segments) to increase the diversity. However, these initializations may not contain the entire object region or include background clutters. To refine each initialized segment, we learn an online SVM model based on color histograms (as used in the clustering stage), and re-estimate the foreground region using an iterative scheme (e.g., one iteration is sufficient in this work) as in the GrabCut method [23].

Tracking Object Segments. Based on multiple initializations from the previous step, we aim to track segments and generate consistent tracklets (as illus-

Fig. 3. An example to track the object under heavy occlusions based on the proposed bi-directional approach with multiple initializations, where initialized segments are denoted as colored rectangular boxes.

trated in Fig. 2). The tracking scheme can better localize objects that may be missed by detection algorithms in a single frame, while maintaining temporal connections between object segments. Since selected segments within the same cluster share similar appearances, we track multiple segments in both forward and backward directions, and group them into two tracklets. Hence, we obtain $2N$ tracklets for each cluster. We note that the bi-directional approach facilitates tracking segments under heavy occlusions (see Fig. 3 for an example). Further note that each initialized segment only tracks a small number of frames until reaching the next initialization, as most tracking methods perform well within a number of frames.

Considering the case of forward tracking from frame $t-1$ to t, the goal is to assign each pixel $x_i^t \in X$ with a foreground or background label $\in \{0, 1\}$. We define an energy function in a Conditional Random Field (CRF):

$$E(X) = U_t(X) + \gamma^s \sum_{(i,j,t)\in\mathcal{N}_t} V_t(x_i^t, x_j^t) + \gamma^t \sum_{(i,j,t)\in\mathcal{N}_t} W_t(x_i^{t-1}, x_j^t), \qquad (1)$$

where U_t is the unary potential to be foreground or background, and V_t and W_t are pairwise potentials for spatial and temporal smoothnesses with weights γ^s and γ^t, respectively. The pairwise terms are defined in a way similar to those in [21]. To reduce the computational load and the effect of background noise, we only segment the object within an estimated object location R_t, obtained as in [29]. Note that we also define $\mathcal{N}_t$ as the neighboring set within this region. For the unary term in (1), we compute appearance and location energies defined by:

$$U_t(X) = \alpha \sum_{(i,t)\in R_t} \Phi_a(x_i^t) + \beta \sum_{(i,t)\in R_t} \Phi_l(x_i^t), \qquad (2)$$

where Φ_a is the appearance term, and Φ_l is the location term. For the appearance term, we learn a SVM model based on color histograms (as used in the clustering stage) from the first frame, and an online SVM model with CNN features [18] updated every frame. The weight α consists of α^{col} and α^{cnn} for the color and CNN features, respectively. By minimizing (1) using the graph cut method [1], we obtain labels and thus the object mask within R_t, and continue to track segments in the next frame.

3.3 Semantic Tracklet Co-selection via Submodular Function

For each video, we generate a set of tracklets where each one is assigned to an object category from the FCN method. However, these tracklets are usually noisy (false negatives) and may not belong to any true object (false positives). In addition, objects within the same category usually share more similarities. To better select object-like tracklets, we collect all those within the same category from all videos to help each other. That is achieved by constructing a graph where the tracklets are nodes, and formulating it as a submodular optimization problem which aims to find a subset that shares more similarities. Once tracklets are selected in each video, we rank different semantic objects based on the submodular energies and find prominent objects.

Graph Construction on Tracklets. We first collect tracklets from all videos where each one is associated with an object category from a set of M categories $\mathcal{L} = \{1, 2, \cdots, M\}$. For each category $l \in \mathcal{L}$, we can find a tracklet set $\mathcal{O}$, and construct a graph $G = (\mathcal{V}, \mathcal{E})$ containing tracklets from all videos (with the same category l), where each node $v \in \mathcal{V}$ is a tracklet and the edges $e \in \mathcal{E}$ model the pairwise relations. For each G, we aim to discover an object-like tracklet set $\mathcal{A}$ of $\mathcal{O}$ by iteratively selecting elements of $\mathcal{O}$ into $\mathcal{A}$.

Submodular Function. Our submodular objective function is designed to find tracklets that meet two criteria: (1) sharing more similarities, (2) maintaining high quality object-like segments. To achieve this, we model the submodular function with a facility location term [13,36] to compute similarities, and a unary term that measures how likely the tracklet belongs to the true object. We first introduce the facility location term defined as:

$$\mathcal{F}(\mathcal{A}) = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{V}} w_{ij} - \sum_{i \in \mathcal{A}} \phi_i, \tag{3}$$

where w_{ij} is the pairwise relation between a potential facility v_i and a node v_j, and ϕ_i is the cost to open a facility fixed to a constant ϵ. In (3), we define w_{ij} as the similarity $S(v_i, v_j)$ to encourage the model to find a similar facility v_i to v_j such that the final selected tracklets share more similarities.

To compute the similarity between two tracklets, we represent each tracklet by a feature vector F_i, and compute their inner product, $S(v_i, v_j) = \langle F_i, F_j \rangle$, as the similarity. For each tracklet, we extract CNN features (same as mentioned in (2)) in each frame and utilize an average pooling method to compute a feature vector that represents each object. Then F_i is computed by averaging feature vectors from all the frames to represent each tracklet. Note that F_i represents appearance of the tracklet in semantics that is learned from CNN, and hence tracklets within the same category are likely to have higher mutual similarities.

However, with only the facility location term, it is not effective in removing all the noisy tracklets in the selected subset $\mathcal{A}$. Hence we propose to include a unary term in the submodular function that can measure the quality of tracklets while preserving the submodularity. The proposed unary term is defined as:

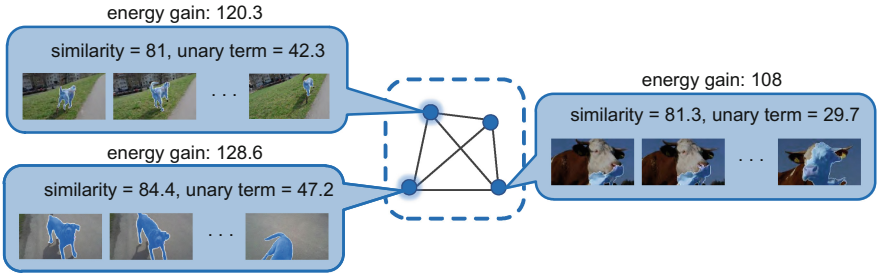

Fig. 4. Illustration of the proposed submodular function for tracklet co-selection. We show three tracklets within the dog category, where the left two tracklets are selected as true objects (denoted as glowing nodes). For each tracklet, we show energy gain, unary term and summed pairwise energy (similarity) in the facility location term. While all three tracklets have high similarity scores, the right tracklet (false positive) has lower energy gain due to low unary term resulting from inconsistent motions and shapes, and hence it is not selected as the object.

$$\mathcal{U}(\mathcal{A}) = \lambda_o \sum_{i \in \mathcal{A}} \Phi_o(i) + \lambda_m \sum_{i \in \mathcal{A}} \Phi_m(i) + \lambda_s \sum_{i \in \mathcal{A}} \Phi_s(i), \tag{4}$$

where $\Phi_o(i)$ measures how likely v_i belongs to the true object (objectness score), and $\Phi_m(i)$ and $\Phi_s(i)$ evaluate the quality of v_i based on the consistency of motions and shapes.

First, we compute $\Phi_o(i) = p_o(i)$ by utilizing probabilities from the FCN output layer according to its category, where $p_o(i)$ is the average probability on all the pixels in v_i. For motion consistency, we use a method similar to [33] and compute motion scores around segment boundaries based on the average gradient magnitude of optical flow estimations [32]. Then we compute $\Phi_m(i)$ by averaging all the motion scores obtained for every two frames. The shape consistency is also considered by computing the intersection-over-union (overlap) ratio between two object segments in adjacent frames. We then compute the variance $\nu_s(i)$ of these overlap ratios, and define $\Phi_s(i) = 1 - \nu_s(i)$, which reflects that larger variance has lower consistency.

Optimization for Tracklet Co-selection. We aim to formulate a submodular function such that tracklets in the selected set $\mathcal{A}$ share more similarities and maintain object-like as well as consistent segments. We combine the facility location term (3) and the unary term (4) with a weight δ into an objective function, and the submodularity is preserved by linearly combining two non-negative terms:

$$\begin{aligned} \max_{\mathcal{A}} \mathcal{C}(\mathcal{A}) = \max_{\mathcal{A}} \mathcal{F}(\mathcal{A}) + \delta \mathcal{U}(\mathcal{A}), \\ \text{s.t. } \mathcal{A} \subseteq \mathcal{O} \subseteq \mathcal{V},\ \mathcal{N}_{\mathcal{A}} \leq \mathcal{N}, \\ \mathcal{H}(\mathcal{A}^i) \geq 0, \\ \mathcal{H}(\mathcal{A}^i) \geq \rho \cdot \mathcal{H}(\mathcal{A}^{i-1}), \end{aligned} \tag{5}$$

Algorithm 1. Tracklet Co-selection for Each Category

Input: $G = (\mathcal{V}, \mathcal{E}), \mathcal{N}, \rho$
Initialization: $\mathcal{A}^0 \leftarrow \emptyset$, $\mathcal{O}^0 \leftarrow \mathcal{V}$, $i \leftarrow 1$
loop
  $a^* = \underset{\{\mathcal{A}^i \in \mathcal{V}\}}{\arg\max}\, \mathcal{H}(\mathcal{A}^i)$, where $\mathcal{A}^i = \mathcal{A}^{i-1} \cup a$
  if $\mathcal{N}_\mathcal{A} > \mathcal{N}$ or $\mathcal{H}(\mathcal{A}^i) < 0$ or $\mathcal{H}(\mathcal{A}^i) < \rho \cdot \mathcal{H}(\mathcal{A}^{i-1})$ when $i \geq 2$ **then**
    break
  end if
  $\mathcal{A}^i \leftarrow \mathcal{A}^{i-1} \cup a^*$, $\mathcal{O}^i \leftarrow \mathcal{O}^{i-1} - a^*$, $i = i + 1$
end loop
Output: $\mathcal{A} \leftarrow \mathcal{A}^i$, $\mathcal{O} \leftarrow \mathcal{O}^i$

where $\mathcal{N}_\mathcal{A}$ is the number of open facilities, and $\mathcal{H}(\mathcal{A}^i)$ is the energy gain at iterations i during iterative optimization, which is defined as: $\mathcal{C}(\mathcal{A}^i) - \mathcal{C}(\mathcal{A}^{i-1})$. We adopt a greedy algorithm to optimize (5) in a way similar to [36]. We start from an empty set of $\mathcal{A}$ and iteratively add an element $a \in \mathcal{V} \backslash \mathcal{A}$ to $\mathcal{A}$ that provides the largest energy gain. The iterative process stops when one of the following conditions is satisfied. First, the number of selected nodes is reached, i.e., $\mathcal{N}_\mathcal{A} > \mathcal{N}$. Second, the energy gain is negative, i.e., $\mathcal{H}(\mathcal{A}^i) < 0$. Third, the ratio of increased energy gain is below a threshold, i.e., $\mathcal{H}(\mathcal{A}^i) < \rho \cdot \mathcal{H}(\mathcal{A}^{i-1})$, when $i \geq 2$. We show the main steps of the tracklet co-selection algorithm for each category l in Algorithm 1 and Fig. 4 illustrates the effectiveness of the proposed submodular function.

After optimizing (5) for each graph G within one category, we select a set of tracklets $\mathcal{T}_l$ for each category l. Considering each video, we can obtain a few tracklets from different sets of $\mathcal{T}_l$, where l can be any category among $\mathcal{L}$. In each video, we then compute the normalized energy gain for each obtained tracklet and re-rank all of them. This is, a normalized gain for a tracklet with category l added at iteration i during optimization is computed as $\mathcal{G}_l^i = \frac{\mathcal{H}(\mathcal{A}^i)}{\mathcal{C}(\mathcal{A}^1)}$, where $\mathcal{C}(\mathcal{A}^1)$ is the energy as the normalization term after adding the first tracklet. Based on the re-ranked results, a threshold (i.e., 0.85 in this work) is applied to all $\mathcal{G}_l^i$ for selecting a set of semantic tracklets that represent prominent objects. To obtain final semantic segmentation results, since object segments from different tracklets may overlap to each other, we choose the one with larger $\mathcal{G}_l^i$ in overlapped regions.

4 Experimental Results

We evaluate the proposed co-segmentation algorithm against the state-of-the-art methods on numerous benchmark datasets. The MATLAB code will be made available at https://sites.google.com/site/yihsuantsai/.

4.1 Experimental Settings

For tracklet generation, we learn an online SVM model with CNN features combining the first three convolutional layers [17] (i.e., 448 dimensional vectors).

For parameters in the graphical model (1) and (2), we use $\alpha^{col} = 1, \alpha^{cnn} = 1, \beta = 0.5, \gamma^s = 3.5$ and $\gamma^t = 1$. In the submodular function, we set ϵ as 3 in the facility location term of (3), and use $\lambda_o = \lambda_m = \lambda_s = 1$ in the unary term of (4). During submodular optimization, we use $\delta = 20$ in (5), and set $\mathcal{N} = 10$ and $\rho = 0.8$ to determine stopping conditions. All these parameters are fixed in the experiments for fair evaluations.

4.2 Youtube-Objects Dataset

The Youtube-Objects dataset [22] contains 10 object categories, and the length of each sequence is up to 400 frames. We evaluate the proposed algorithm in a subset of 126 videos with more than 20000 frames, where the pixel-wise annotations in every 10 frames are provided by [10]. Note that, different from previous video co-segmentation datasets [3,34], appearances and shapes of objects from the same category in this dataset are significantly different.

We first conduct experiments in a weakly supervised manner, where a semantic label is given for each video. Next, we evaluate our algorithm in a way that object categories are unknown in videos. Table 1 shows segmentation results of the proposed method and other state-of-the-art approaches. We use the intersection-over-union (overlap) ratio to evaluate all the methods.

Weakly Labeled Videos. For the video labeled with a semantic category, we use FCN segments belonging to its video-level category as initializations, such that tracklets generated in each video (as described in Sect. 3.2) are all associated with the same category. We compare our approach with other supervised tracking-based [7,20] or weakly supervised [35][1] methods. Table 1 shows that the proposed method with weak supervision performs favorably in terms of overlap ratio, especially in 7 out of 10 categories.

In general, our method performs well on non-rigid objects (*bird*, *cat*, *dog*, *horse*) and fast moving objects (*car*, *train*). As the appearances and shapes of these objects vary significantly, it is challenging to segment these objects from all videos accurately. Although the recent method [35] utilizes object detectors and generates proposals to localize objects in each frame, it is less effective for videos with large appearance and shape variations as the generated proposals are usually noisy and less consistent across videos. In contrast, the proposed tracking-based algorithm is able to capture detailed appearance and shape changes, and hence generate tracklets consistently for segmentation.

Semantic Co-segmentation. In addition to weakly supervised settings, the proposed algorithm can segment objects and discover the corresponding object categories without any supervision. Table 1 shows our segmentation results compared with the state-of-the-art unsupervised method [21]. The proposed algorithm generates more accurate segmentation results in most categories with significant improvement (e.g., more than 10 % gain in *boat, cat* and *train*).

[1] [35] evaluates the method on their annotated images, and we obtain their results on the same annotation set [10] directly from the authors.

Table 1. Segmentation results on the Youtube-Objects dataset with the overlap ratio.

Category	[7]	[35]	Ours	[20]	[21]	Baseline (FCN)	Ours
Supervised?	Y	weakly	weakly	N	N	N	N
Aeroplane	**73.6**	72.4	69.3	13.7	**70.9**	60.8	69.3
Bird	56.1	66.6	**76.1**	12.2	70.6	69.7	**76.0**
Boat	**57.8**	43.0	57.2	10.8	42.5	44.7	**53.5**
Car	33.9	58.9	**70.4**	23.7	65.2	60.3	**70.4**
Cat	30.5	36.4	**67.7**	18.6	52.1	53.9	**66.8**
Cow	41.8	58.2	**59.7**	16.3	44.5	**52.8**	49.0
Dog	36.8	48.7	**64.2**	18.0	**65.3**	52.8	47.5
Horse	44.3	49.6	**57.1**	11.5	53.5	42.4	**55.7**
Motorbike	**48.9**	41.4	44.1	10.6	44.2	**47.3**	39.5
Train	39.2	49.3	**57.9**	19.6	29.6	**54.7**	53.4
Mean	46.3	52.4	**62.3**	15.5	53.8	53.9	**58.1**

It demonstrates the effectiveness of our co-segmentation scheme that links relations between semantic objects from all videos, which is not addressed in [21].

To evaluate the effectiveness of the proposed tracking-based algorithm for tracklet generation, we establish a baseline method which directly groups FCN segments from every frame into a tracklet for each category (i.e., without using tracking). We then use the same submodular function for tracklet co-selection (Sect. 3.3). Compared to this baseline method, the proposed algorithm performs well on most categories, especially for deformable objects such as *bird*, *cat* and *horse*, as consistent tracklets can be extracted. However, the proposed algorithm does not perform well in some videos (*cow*, *motorbike*) as some segments are not initialized well, which causes inaccurate tracking results in these videos.

Compared to the proposed algorithm with weakly supervised setting, the results on categories such as *aeroplane*, *bird* and *car* have identical and high overlap ratios. It shows that without providing video-level labels, our co-segmentation approach can reduce noisy segments that are generated from other false categories, and hence retain high accuracies as with weakly supervised setting. Moreover, it is worth noticing that the proposed algorithm without supervision, already performs favorably against the state-of-the-art method that requires weak supervision [35].

Different from other methods [21,35], the proposed algorithm can segment objects as well as discover object categories (labels). We evaluate the classification accuracy for predicting object categories based on ranked tracklets, and the average precision (AP) is 85.3 on average over all categories. The results show that with the proposed submodular function and re-ranking in each video, false positives can be reduced, and hence prominent objects are discovered. We show qualitative results in Fig. 5, and more results are presented in the supplementary material.

Fig. 5. Example results for semantic co-segmentation on the Youtube-Objects dataset (without knowing object categories). The colors overlapping on the objects indicate different semantic labels. The results show that our method is able to track and segment (multiple) objects under challenges such as occlusions, fast movements, deformed shapes, scale changes and cluttered backgrounds. Best viewed in color with enlarged images. (Color figure online)

4.3 MOViCS Dataset

The MOViCS dataset [3], which contains 4 sets with 11 sequences, is used for evaluation on multi-class video co-segmentation. For each set, at least one common object appears in all videos, while the number of object categories is unknown. The proposed algorithm is evaluated against three state-of-the-art methods including image co-segmentation (ICS) [11], video co-segmentation (VCS) [3] and RMWC [34]. We use the unsupervised method [21] as a baseline and produce segments in each frame as initializations for tracklet generation (Sect. 3.2). In addition, since categories are not known for different segments at this stage, one graph including tracklets from all videos is constructed for co-selecting tracklets in each video.

Based on the evaluation metric in [3], Table 2 shows that the proposed algorithm performs well in all the video sets, especially in the *tiger* set. As the variations of objects in some videos are large, other approaches are less effective in segmenting objects in these videos. In contrast, our method works for objects with various appearances in different videos by utilizing the submodular optimization that accounts for appearances, shapes and motions together to co-select tracklets containing common objects. We show qualitative comparisons to other methods in Fig. 6.

4.4 Safari Dataset

In addition to co-segmentation in videos where each set contains at least one common object, our method is able to segment objects given a collection of sequences without any prior knowledge. The Safari dataset [34] contains 9 videos with 5

Fig. 6. Example results for object co-segmentation on the MOViCS dataset. Segmentation outputs are indicated as colored contours, where each color represents an instance. Compared to the state-of-the-art approach [34] and the baseline method [21] that often produce noisy segments or missing objects, our method obtains better segmentation results. Best viewed in color. (Color figure online)

Table 2. Segmentation results on the MOViCS dataset with the overlap ratio.

Video Set	ICS [11]	RMWC [34]	VCS [3]	Baseline [21]	Ours
Chicken & Turtle	8.0	86.0	65.0	73.6	**87.7**
Zebra & Lion	23.0	58.8	48.0	45.9	**71.3**
Giraffe & Elephant	7.0	52.8	52.0	36.5	**59.0**
Tiger	30.0	33.6	30.0	44.1	**70.9**
Mean	17.0	57.8	48.8	50.0	**72.2**

object categories, where each video may contain one or two object categories. To evaluate the proposed algorithm, we input these 9 videos together and segment common objects. Note that, we use [21] as the baseline method for single video object segmentation. Then we initialize these segments to generate tracklets and construct a graph for tracklet co-selection.

Table 3 shows the results by the proposed algorithm and two state-of-the-art methods. In 4 out of 5 categories, our method achieves better results over the other methods. The VCS [3] method is not effective for the general setting when videos contain unknown types of object categories, and hence generates less accurate results. The RMWC method [34] relies on object proposals and does not generate consistent tracklets across videos when more than one object category is involved. In our proposed algorithm, we utilize tracking-based method to generate consistent tracklets, and segment objects via submodular optimization

Fig. 7. Example results for object co-segmentation on the Safari dataset. Segmentation outputs are indicated as colored contours, where each color represents an instance. Compared to the state-of-the-art approach [34] (second row) and the baseline method [21] (first row) that often produce noisy segments, false positives or missing objects, our method obtains better segmentation results. Best viewed in color.

Table 3. Segmentation results on the Safari dataset with the overlap ratio.

Object	RMWC [34]	VCS [3]	Baseline [21]	Ours
Buffalo	86.9	68.6	90.0	**91.3**
Elephant	35.3	26.6	73.8	**74.9**
Giraffe	2.4	2.4	9.8	**15.8**
Lion	**31.7**	30.2	19.0	21.9
Sheep	36.3	4.8	32.3	**65.8**
Mean	38.5	26.5	45.0	**54.0**

in multiple videos without any assumption on the commonality of objects in the videos. We show some example results in Fig. 7.

5 Concluding Remarks

In this paper, we present a novel algorithm to segment objects and understand their visual semantics from a collection of videos. To exploit semantic information, we first assign a category for each discovered segment in videos via the FCN method. A tracking-based approach is presented to generate consistent tracklets across videos. We then link the relations between videos by constructing graphs which contain tracklets from different videos. Without any assumption of objects appearing in videos, we formulate a submodular optimization problem and co-select tracklets, which accounts for their appearances, shapes and motions. This step considers other sequences and reduces noisy tracklets that can not be filtered out within a single video. As a result, prominent objects are discovered and segmented in videos. Extensive experimental results on the Youtube-Objects, MOViCS and Safari datasets show that our method performs favorably against the state-of-the-art approaches in terms of visual quality and accuracy.

Acknowledgments. This work is supported in part by the NSF CAREER grant #1149783, NSF IIS grant #1152576, and gifts from Adobe and Nvidia. G. Zhong is sponsored by China Scholarship Council.

References

1. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. PAMI **26**(9), 1124–1137 (2004)
2. Chen, X., Shrivastava, A., Gupta, A.: Enriching visual knowledge bases via object discovery and segmentation. In: CVPR (2014)
3. Chiu, W.C., Fritz, M.: Multi-class video co-segmentation with a generative multi-video model. In: CVPR (2013)
4. Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: CVPR (2015)
5. Everingham, M., Gool, L.J.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)
6. Fu, H., Xu, D., Zhang, B., Lin, S.: Object-based multiple foreground video co-segmentation. In: CVPR (2014)
7. Godec, M., Roth, P.M., Bischof, H.: Hough-based tracking of non-rigid objects. In: ICCV (2011)
8. Guo, J., Cheong, L.-F., Tan, R.T., Zhou, S.Z.: Consistent foreground co-segmentation. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9006, pp. 241–257. Springer, Heidelberg (2015)
9. Hartmann, G., Grundmann, M., Hoffman, J., Tsai, D., Kwatra, V., Madani, O., Vijayanarasimhan, S., Essa, I., Rehg, J., Sukthankar, R.: Weakly supervised learning of object segmentations from web-scale video. In: ECCV Workshop (2012)
10. Jain, S.D., Grauman, K.: Supervoxel-consistent foreground propagation in video. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 656–671. Springer, Heidelberg (2014)
11. Joulin, A., Bach, F., Ponce, J.: Multi-class cosegmentation. In: CVPR (2012)
12. Kwak, S., Cho, M., Laptev, I., Ponce, J., Schmid, C.: Unsupervised object discovery and tracking in video collections. In: ICCV (2015)
13. Lazic, N., Givoni, I., Frey, B., Aarabi, P.: Floss: Facility location for subspace segmentation. In: ICCV (2009)
14. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV (2011)
15. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV (2013)
16. Liu, X., Tao, D., Song, M., Ruan, Y., Chen, C., Bu, J.: Weakly supervised multi-class video segmentation. In: CVPR (2014)
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
18. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: ICCV (2015)
19. Nagaraja, N.S., Schmidt, F., Brox, T.: Video segmentation with just a few strokes. In: ICCV (2015)
20. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. PAMI **36**(6), 1187–1200 (2014)

21. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV (2013)
22. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR (2012)
23. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: SIGGRAPH (2004)
24. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: CVPR (2013)
25. Rubio, J.C., Serrat, J., Antonio, L., Paragios, N.: Unsupervised co-segmentation through region matching. In: CVPR (2012)
26. Rubio, J.C., Serrat, J., López, A.: Video co-segmentation. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part II. LNCS, vol. 7725, pp. 13–24. Springer, Heidelberg (2013)
27. Tang, K., Joulin, A., Li, L.J., Fei-Fei, L.: Co-localization in real-world images. In: CVPR (2014)
28. Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: CVPR (2013)
29. Tsai, Y.H., Yang, M.H., Black, M.J.: Video segmentation via object flow. In: CVPR (2016)
30. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: CVPR (2011)
31. Wang, L., Hua, G., Sukthankar, R., Xue, J., Zheng, N.: Video object discovery and co-segmentation with extremely weak supervision. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 640–655. Springer, Heidelberg (2014)
32. Wulff, J., Black, M.J.: Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In: CVPR (2015)
33. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: CVPR (2013)
34. Zhang, D., Javed, O., Shah, M.: Video object co-segmentation by regulated maximum weight cliques. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 551–566. Springer, Heidelberg (2014)
35. Zhang, Y., Chen, X., Li, J., Wang, C., Xia, C.: Semantic object segmentation via detection in weakly labeled video. In: CVPR (2015)
36. Zhu, F., Jiang, Z., Shao, L.: Submodular object recognition. In: CVPR (2014)