Video Object Discovery and Co-segmentation with Extremely Weak Supervision

Le Wang[♭⋆], Gang Hua[♮], Rahul Sukthankar[♯], Jianru Xue[♭], and Nanning Zheng[♭]

[♭]Xi'an Jiaotong University [♮]Stevens Institute of Technology [♯]Google Research

Abstract. Video object co-segmentation refers to the problem of simultaneously segmenting a common category of objects from multiple videos. Most existing video co-segmentation methods assume that all frames from all videos contain the target objects. Unfortunately, this assumption is rarely true in practice, particularly for large video sets, and existing methods perform poorly when the assumption is violated. Hence, any practical video object co-segmentation algorithm needs to identify the relevant frames containing the target object from all videos, and then co-segment the object only from these relevant frames. We present a spatiotemporal energy minimization formulation for simultaneous video object discovery and co-segmentation across multiple videos. Our formulation incorporates a spatiotemporal auto-context model, which is combined with appearance modeling for superpixel labeling. The superpixel-level labels are propagated to the frame level through a multiple instance boosting algorithm with spatial reasoning (Spatial-MILBoosting), based on which frames containing the video object are identified. Our method only needs to be bootstrapped with the frame-level labels for a few video frames (*e.g.*, usually 1 to 3) to indicate if they contain the target objects or not. Experiments on three datasets validate the efficacy of our proposed method, which compares favorably with the state-of-the-art.

Keywords: video object discovery, video object co-segmentation, spatiotemporal auto-context model, Spatial-MILBoosting.

1 Introduction

The problem of simultaneously segmenting a common category of objects from two or more videos is known as video object co-segmentation. Compared with object segmentation from a single image, the benefit is that the appearance and/or structure information of the target objects across the videos are leveraged for segmentation. Several previous methods [9,13,27] have attempted to harness such information for video object co-segmentation.

However, these methods [9, 13, 27] all made the assumption that all frames from all videos contain the target object, *i.e.*, all frames are relevant. Moreover, a closer look at the video datasets employed in previous papers reveals that the object instances in different videos are frequently the same object [9], or only exhibit small variations in color, shape, pose, size, and location [13, 27]. These limitations render such methods

⋆ Le Wang participated in this project while working at Stevens Institute of Technology as a visiting Ph.D. student supervised by Prof. Gang Hua.

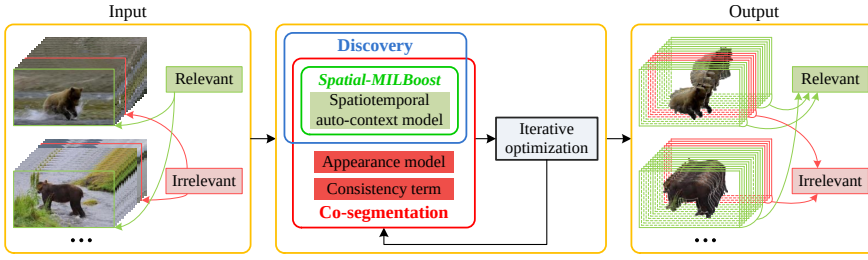

Fig. 1. The flowchart of our video object discovery and co-segmentation method.

less applicable to real-world videos, such as those online videos gathered from a search engine in response to a specific query. The common objects in these videos are usually just of the same category, exhibiting dramatic variations in color, size, shape, pose, and viewpoint. Moreover, it is not uncommon for such videos to contain many irrelevant frames where the target objects are not present. This suggests that a practical video object co-segmentation method should also be capable of identifying the frames that contain the objects, *i.e.*, discover the objects.

We present a spatiotemporal energy minimization formulation to simultaneously discover and co-segment the target objects from multiple videos containing irrelevant frames. Fig. 1 presents the flowchart of our method. Bootstrapped from just a few (often 1-3) labeled frames indicating whether they are relevant or not, our method incurs a top-down modeling to propagate the frame-level label to the superpixels through a multiple instance boosting algorithm with spatial reasoning, namely Spatial-MILBoosting. From bottom up, the labels of the superpixels are jointly determined by a spatiotemporal auto-context model induced from the Spatial-MILBoosting algorithm and an appearance model using colors.

The learning of the spatiotemporal auto-context model, cast together with the color based appearance model as the data term, is embedded in a spatiotemporal energy minimization framework for joint object discovery and co-segmentation. Due to the embedded formulation, the learning of the spatiotemporal auto-context model (hence the object discovery), and the minimization of the energy function conducted by min-cut [6,7] (hence the object co-segmentation), are performed iteratively until convergence. The final output of our method includes a frame-level label for each frame indicating if it contains the target object, and a superpixel-level labeling of the target object for each identified relevant frame.

As a key component of our formulation, our proposed spatiotemporal auto-context model extends the original auto-context model [31] to also capture the temporal context. Our embedded formulation also facilitates learning the model with only weak supervision with frame-level labels using the Spatial-MILBoosting algorithm. The Spatial-MILBoosting allows information to be propagated between the frame level and the superpixel level, and hence facilitates the discovery of the objects and the co-segmentation by effectively exploiting the spatiotemporal context across multiple videos.

In summary, the key contributions of this paper are: (1) We propose a method to simultaneously discover and co-segment of a common category of objects from multiple videos containing irrelevant frames. (2) To facilitate both the object discovery and co-segmentation, we model the spatiotemporal contextual information across multiple videos by a spatiotemporal auto-context model learned from a Spatial-MILBoosting algorithm. (3) To exactly evaluate the proposed method, we collect and release a new 10-categories video object co-segmentation dataset with ground truth frame-level labels for all frames and pixel-wise segmentation labels for all relevant frames.

2 Related Work

Video object discovery. Video object discovery has recently been extensively studied, in both unsupervised [18,42] or weakly supervised [19,24] settings. Liu and Chen [18] proposed a latent topic model for unsupervised object discovery in videos by combining pLSA with Probabilistic Data Association filter. Zhao *et al.* [42] proposed a topic model by incorporating a word co-occurrence prior into LDA for efficient discovery of topical video objects from a set of key frames. Liu *et al.* [19] engaged human in the loop to provide a few labels at the frame level to roughly indicate the main object of interest. Prest *et al.* [24] proposed a fully automatic method to learn a class-specific object detector from weakly annotated real-world videos. Tuytelaars *et al.* [32] surveyed the unsupervised object discovery methods, but with the focus on still images. In contrast, our video object discovery is achieved by propagating superpixel-level labels to frame level through a Spatial-MILBoosting algorithm.

Video object segmentation/co-segmentation. Video object segmentation refers to the task of separating the objects from the background in a video, either interactively [4,28, 30] or automatically [8,12,16,17,20,22,23,41]. A number of methods have focused on finding the object-like proposals for this problem [16,20,23,41]. Several methods track feature points or local regions over frames, and then cluster the resulting tracks based on pairwise [8,30] or triplet similarity measures [17,22]. Tang *et al.* [28] proposed an algorithm for annotating spatiotemporal segments based on video-level labels. Grundmann *et al.* [12] cluster a video into spatiotemporal consistent supervoxels.

Several video object co-segmentation methods [9, 13, 27] have been proposed recently to simultaneously segment a common category of objects from two or more videos. They made the assumption that all frames from all videos should contain the target object. Chiu and Fritz [10] proposed an algorithm to conduct multi-class video object co-segmentation, in which the number of object classes and the number of instances are unknown in each frame and video. Our method jointly discovers and co-segments the target objects from multiple videos, in which an unknown number of frames do not contain the target objects at all.

Image co-segmentation. Our work is also related to image co-segmentation [5, 11, 15, 26, 33, 34], where the appearance or structure consistency of the foreground objects across the image collection is exploited to benefit object segmentation. The objective of image co-segmentation is to jointly segment a specific object from two or more images, and it is assumed that all images contain that object. There are also several co-segmentation methods that conduct the co-segmentation of noisy image collections [25,

Table 1. Principal notations.

$\mathcal{V}$	A collection of N videos	l_i^n	The label of f_i^n, $l_i^n \in \{0, 1\}$, where 1
$\mathcal{L}$	The frame-level labels of $\mathcal{V}$		means that f_i^n is relevant, *i.e.,* f_i^n contains
$\mathcal{B}$	A segmentation of $\mathcal{V}$		the target object
V^n	The nth video in $\mathcal{V}$ with N^n frames	b_i^n	A segmentation of f_i^n
L^n	The frame-level labels of V^n	s_{ij}^n	The jth superpixel in f_i^n
B^n	A segmentation of V^n	b_{ij}^n	The label of s_{ij}^n, $b_{ij}^n \in \{0, 1\}$, where 1
f_i^n	The ith frame of V^n with N_i^n superpixels		means that s_{ij}^n belongs to the target object

38], in which several images do not contain the target objects. In our work, we focus on video object discovery and co-segmentation with noisy video collections, where many frames may not contain the target objects.

3 Problem Formulation

For ease of presentation, we first summarize the main notations in Table 1. Then we present the proposed spatiotemporal energy minimization framework for simultaneous object discovery and co-segmentation across multiple videos, along with details of the spatiotemporal context model and the Spatial-MILBoosting algorithm.

Given a set of videos $\mathcal{V}$, our objective is to obtain a frame-level label l_i^n for each frame f_i^n indicating if it is a relevant frame that contains the target objects, and a superpixel-level labeling b_i^n of the target object for each identified relevant frame f_i^n ($l_i^n = 1$). We cast this problem into a spatiotemporal energy minimization framework. Then, our energy function for simultaneous object discovery and co-segmentation from multiple videos $\mathcal{V}$ becomes

$$\begin{aligned} E(\mathcal{B}) = & \sum_{s_{ij}^n \in \mathcal{V}} D_j^1(b_{ij}^n) + \sum_{s_{ij}^n \in V^n} D_j^2(b_{ij}^n) \\ & + \sum_{s_{ij}^n, s_{ik}^n \in \mathcal{N}_j} S_{jk}^1(b_{ij}^n, b_{ik}^n) + \sum_{s_{ij}^n, s_{uk}^n \in \bar{\mathcal{N}}_j} S_{jk}^2(b_{ij}^n, b_{uk}^n), \\ & n = 1, \ldots, N, i = 1, \ldots, N^n, j = 1, \ldots, N_i^n, \end{aligned} \tag{1}$$

where $D_j^1(b_{ij}^n)$ and $D_j^2(b_{ij}^n)$ compose the data term, measuring the cost of labeling superpixel s_{ij}^n to be b_{ij}^n from a spatiotemporal auto-context model and a color based appearance model, respectively. The spatiotemporal auto-context model builds a multi-layer Boosting classifier on context features surrounding a superpixel to predict if it is associated with the target concept, where subsequent layer is working on the probability maps from the previous layer, detailed below in Sec. 3.1. Hence, $D_j^1(b_{ij}^n)$ relies on the discriminative probability maps estimated by a learned spatiotemporal auto-context model. It is learned to model the spatiotemporal contextual information across multiple videos $\mathcal{V}$, and thus is video independent. While the appearance model is estimated by capturing the color distributions of the target objects and the backgrounds for each video V^n, and thus is video dependent.

$S^1_{jk}(b^n_{ij}, b^n_{ik})$ and $S^2_{jk}(b^n_{ij}, b^n_{uk})$ compose the consistency term, constraining the segmentation labels to be both spatially and temporally consistent. $\mathcal{N}_j$ is the spatial neighborhood of s^n_{ij} in f^n_i. $\bar{\mathcal{N}}_j = \{\overleftarrow{s}^n_{ij}, \overrightarrow{s}^n_{ij}\}$ is the temporal neighborhood of s^n_{ij}, *i.e.*, its corresponding next superpixel $\overrightarrow{s}^n_{ij}$ in f^n_{i+1} and previous superpixel $\overleftarrow{s}^n_{ij}$ in f^n_{i-1}. The superpixels are computed by using SLIC [1], due to its superiority in terms of adherence to boundaries, as well as computational and memory efficiency. However, the proposed method is not tied to any specific superpixel method, and one can choose others.

The particular spatiotemporal auto-context model embedded in the energy function is learned through a multiple instance learning algorithm with spatial reasoning (Spatial-MILBoosting), and hence it can propagate information between the frame level and the superpixel level. From top down, the label of frame is propagated to the superpixel level to facilitate the energy minimization for co-segmentation; from bottom up, the labels of superpixels are propagated to the frame level to identify which frame is relevant. Bootstrapped from just a few frame-level labels, the learning of the spatiotemporal auto-context model (hence the object discovery), and the minimization of the energy function conducted by min-cut [6, 7] (hence the object co-segmentation) are performed iteratively until it converges. At each iteration, the spatiotemporal auto-context model, the appearance model, and the consistency term are updated based on the new segmentation $\mathcal{B}$ of $\mathcal{V}$.

3.1 Spatiotemporal Auto-context Model

We extend the auto-context model originally proposed by Tu [31] and later tailored by Wang *et al.* [36, 37, 40] for video object discovery and co-segmentation. The original auto-context model builds a multi-layer Boosting classifier on image and context features surrounding a pixel to predict if it is associated with the target concept, where subsequent layer is working on the probability maps from the previous layer. In previous works, it just modeled the spatial contextual information, either from a single image [36, 40], or a set of labeled [31] or unlabeled [37,38] images. Here, we extend it to capture both the spatial and temporal contextual information across multiple videos, and the extended model operates on superpixels instead of pixels.

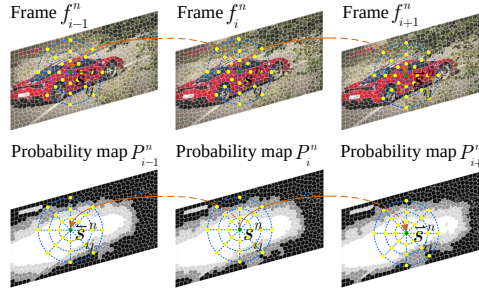

Fig. 2. The spatiotemporal auto-context feature.

Spatiotemporal auto-context feature. Let $\mathbf{c}^n_{ij}$ denote the context feature of superpixel s^n_{ij}, $P^n \in \mathcal{P}$ the probability map set for video V^n, P^n_i the probability map for frame f^n_i, p^n_{ij} the probability value of superpixel s^n_{ij}. The sampling structure of the spatiotemporal auto-context model on the discriminative probability maps are illustrated in Fig. 2. $\mathbf{c}^n_{ij}$

consists of a backward-frame part, a current-frame part and a forward-frame part as

$$\mathbf{c}_{ij}^n = \{\{\overleftarrow{p}_{ij}^n(k)\}, \{p_{ij}^n(k)\}, \{\overrightarrow{p}_{ij}^n(k)\}\}_{k=1}^{N_c}, \tag{2}$$

where $p_{ij}^n(k)$, $\overleftarrow{p}_{ij}^n(k)$ and $\overrightarrow{p}_{ij}^n(k)$ are the probability values of the kth point on the sampling structure centered at s_{ij}^n in P_i^n, its corresponding previous superpixel $\overleftarrow{s}_{ij}^n$ in P_{i-1}^n, and its corresponding next superpixel $\overrightarrow{s}_{ij}^n$ in P_{i+1}^n, respectively. N_c is the number of sampled points on the sampling structure for the current superpixel in each frame, and it is set to be 41 in our experiments. Here, we find the corresponding previous and next superpixels of current superpixel between neighboring frames using optical flow [39]. If the pixel number of the intersection between a superpixel in the current frame and its corresponding superpixel in neighboring frames, identified from the optical flow vector displacements of current superpixel, is larger than half of the pixel number of the current superpixel, it is selected as the temporal neighbor.

Update the spatiotemporal auto-context classifier. In the first round of the iterative learning of the spatiotemporal auto-context model, the training set is built as

$$\mathbf{S}_1 = \{\{\mathbf{C}_{i'}^n(\alpha), l_{i'}^n(\alpha)\} | n = 1, \ldots, N; i' = 1', \ldots, N^{n\prime}; \alpha = 0, 1\}, \tag{3}$$

where i' is the index of frame $f_{i'}^n$ that was manually labeled by the user as relevant ($l_{i'}^n = 1$) or irrelevant ($l_{i'}^n = 0$). $N^{n\prime}$ is the number of labeled frames in video V^n, and it is set to be 1 to 3 in our experiments. $\mathbf{C}_{i'}^n = \{\mathbf{c}_{i'j}^n\}_{j=1}^{N_{i'}^n}$ are the context features of superpixels in $f_{i'}^n$, and $\mathbf{C}_{i'}^n(\alpha)$ are the context features in the object ($\alpha = 1$) or background ($\alpha = 0$) of $f_{i'}^n$. We treat $\mathbf{C}_{i'}^n(\alpha)$ as a *bag*, and $\mathbf{c}_{i'j}^n$ as an *instance*. $l_{i'}^n(\alpha)$ is the label of bag $\mathbf{C}_{i'}^n(\alpha)$, and it equals to 1 when both $l_{i'}^n$ and α equal to 1, and 0 otherwise. In other words, we treat the objects of the relevant frames as positive bags, the backgrounds of the relevant frames and both the objects and backgrounds of the irrelevant frames as negative bags. The initial segmentations $\mathcal{B}$ for $\mathcal{V}$ are obtained by using an objectness measure [2] and a saliency measure [14], and the probability maps $\mathcal{P}$ for $\mathcal{V}$ are initialized by averaging the scores returned by objectness and saliency.

Then, the first classifier $H(\cdot)$ is learned on $\mathbf{S}_1$ using Spatial-MILBoosting, detailed immediately below. We proceed to use the learned classifier to classify all the context features of the objects and backgrounds of all frames in $\mathcal{V}$, and obtain the new probability map set $\mathcal{P}$ for $\mathcal{V}$, where the new probability of superpixel s_{ij}^n being positive is given by the learned classifier as

$$p_{ij}^n = \frac{1}{1 + \exp\left(-H(\mathbf{c}_{ij}^n)\right)}. \tag{4}$$

The data term based on the spatiotemporal auto-context model in Eq(1) is defined as

$$D_j^1(b_{ij}^n) = -\log p_{ij}^n. \tag{5}$$

The probability of the object or background (*bag*) of frame f_i^n being positive is a "Noisy OR" defined as

$$p_i^n(\alpha) = 1 - \prod_{j=1}^{N_i^n(\alpha)} (1 - p_{ij}^n), \tag{6}$$

where $N_i^n(\alpha)$ denotes the number of superpixels (*instances*) in the object or background (*bag*) of frame f_i^n. In this way, the trained auto-context classifier can propagate

Algorithm 1. Spatial-MILBoosting - Training

Input: Training set $\{\mathbf{x}_i, l_i\}_{i=1}^{N}$ of N bags, where each bag $\mathbf{x}_i = \{x_{ij}\}_{j=1}^{N_i}$ containing N_i instances, the bag label $l_i \in \{0, 1\}$.

1. Initialize the instance weights $w_{ij} = 2 * (l_i - 0.5)$ and the instance classifier $H = 0$
2. Initialize estimated margins $\{\hat{y}_{ij}\}_{i,j=1}^{N,N_i}$ to 0
3. For $t = 1, \ldots, T$
   a. Set $\bar{x}_{ij} = \{\hat{y}_{ik} | x_{ik} \in \text{Nbr}(x_{ij})\}$
   b. Train weak *data* classifier h_t^d on the data $\{x_{ij}, l_i\}_{i,j=1}^{N,N_i}$ and the weights $\{\omega_{ij}\}_{i,j=1}^{N,N_i}$ as $h_t^d(x_{ij}) = \arg\max_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(x_{ij}) w_{ij}$
   c. Train weak *spatial* classifier h_t^s on the data $\{\bar{x}_{ij}, l_i\}_{i,j=1}^{N,N_i}$ and the weights $\{\omega_{ij}\}_{i,j=1}^{N,N_i}$ as $h_t^s(\bar{x}_{ij}) = \arg\max_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(\bar{x}_{ij}) w_{ij}$
   d. Set $\epsilon^d = \sum_{i,j} \omega_{ij} |h_t^d(x_{ij}) - l_i|$ and $\epsilon^s = \sum_{i,j} \omega_{ij} |h_t^s(\bar{x}_{ij}) - l_i|$
   e. Set $h_t(x_{ij}) = \begin{cases} h_t^d(x_{ij}) \text{ if } \quad \epsilon^d < \epsilon^s \\ h_t^s(\bar{x}_{ij}) \quad \text{otherwise} \end{cases}$
   f. Find λ_t via line search to minimize likelihood $L(H) = \prod_i (q_i)^{l_i} (1 - q_i)^{(1-l_i)}$ as $\lambda_t = \arg\max_\lambda L(H + \lambda h_t)$
   g. Update margins $\hat{y}_{ij}$ to be $\hat{y}_{ij} = H(x_{ij}) = \hat{y}_{ij} + \lambda_t h_t(x_{ij})$
   h. Compute the instance probability $q_{ij} = \frac{1}{1+\exp(-\hat{y}_{ij})}$
   i. Compute the bag probability $q_i = 1 - \prod_{j=1}^{N_i} (1 - q_{ij})$
   j. Update the instance weights $w_{ij} = \frac{\partial \log L(H)}{\partial y_{ij}} = \frac{l_i - q_i}{q_i} q_{ij}$

Output: Instance classifier $H(x_{ij}) = \sum_{t=1}^{T} \lambda_t h_t(x_{ij})$.

superpixel-level labels indicating if the superpixels belong to the target objects to the object (or background) level label indicating if it contains the target object.

From the second round of the iterative learning process, we update the training set as

$$\mathbf{S}_2 = \{\{\mathbf{C}_i^n(\alpha), l_i^n(\alpha)\} | n = 1, \ldots, N; i = 1, \ldots, N^n; \alpha = 0, 1\}, \tag{7}$$

and learn a new classifier on the updated context features, which are based on the discriminative probability map set $\mathcal{P}$ obtained from the previous iteration. Then, the new $\mathcal{P}$ for $\mathcal{V}$ are computed by the new spatiotemporal auto-context classifier. This process will iterate until convergence, where $\mathcal{P}$ no longer changes. Indeed, the spatiotemporal auto-context model is alternatively updated with the iterative co-segmentation of $\mathcal{V}$, *i.e.*, the iterative minimization of the energy in Eq(1).

Spatial-MILBoosting algorithm. Compared to the original MILBoost algorithm [35], we incorporate the spatial information between the neighboring superpixels [3] into the multiple instance boosting algorithm [19, 35] to infer whether the superpixel is positive or not, and name this algorithm Spatial-MILBoosting. To present the algorithm in a more general sense, we use $\mathbf{x}_i$, l_i and $x_{ij} \in \mathbf{x}_i$ instead of $\mathbf{C}_i^n(\alpha)$, $l_i^n(\alpha)$ and $\mathbf{c}_{ij}^n \in \mathbf{C}_i^n(\alpha)$ to denote the *bag*, its *label* and its *instance*, respectively. The training and testing details of Spatial-MILBoosting are presented in Alg. 1 and Alg. 2, respectively.

The score of the instance x_{ij} is $y_{ij} = H(x_{ij})$, where $H(x_{ij}) = \sum_{t=1}^{T} \lambda_t h_t(x_{ij})$ is a weighted sum of weak classifiers. The probability of the instance x_{ij} being positive is

Algorithm 2. Spatial-MILBoosting - Testing

Input: Unlabeled testing set $\{x_{ij}\}_{i,j=1}^{N,N_i}$, and the instance classifier $H(\cdot)$.

1. Initialize estimated margins $\{\hat{y}_{ij}\}_{i,j=1}^{N,N_i}$ to 0
2. For $t = 1, \ldots, T$
   a. Set $\bar{x}_{ij} = \{\hat{y}_{ik} | x_{ik} \in \text{Nbr}(x_{ij})\}$
   b. Update margins $\hat{y}_{ij}$ to be $\hat{y}_{ij} + \lambda_t h_t(x_{ij})$

Output: Labels $\{\hat{y}_{ij}\}_{i,j=1}^{N,N_i}$.

defined as a standard logistic function,

$$q_{ij} = \frac{1}{1 + \exp(-y_{ij})}. \tag{8}$$

The probability of the bag $\mathbf{x}_i$ being positive is a "Noisy OR" as

$$q_i = 1 - \prod_{j=1}^{N_i} (1 - q_{ij}). \tag{9}$$

The goal now is to estimate λ_t and h_t, so q_{ij} approaches its true value. The likelihood assigned to a set of training bags is $L(H) = \prod_i (q_i)^{l_i}(1 - q_i)^{(1-l_i)}$, and is maximum when $q_i = l_i$, where $l_i \in \{0, 1\}$ is the label of bag $\mathbf{x}_i$. To find an instance classifier that maximizes the likelihood, we compute the derivative of the log-likelihood with respect to y_{ij} as $\frac{\partial \log L(H)}{\partial y_{ij}} = w_{ij} = \frac{l_i - q_i}{q_i} q_{ij}$.

In each round t of gradient descent, one solves the optimal weak *instance* classifier $h_t(\cdot)$. Here, we train a weak *data* classifier on the data $\{x_{ij}, l_i\}_{i,j=1}^{N,N_i}$ and the weights $\{\omega_{ij}\}_{i,j=1}^{N,N_i}$ as $h_t^d(x_{ij}) = \arg\max_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(x_{ij}) w_{ij}$. Meanwhile, we train a weak *spatial* classifier on the data $\{\bar{x}_{ij}, l_i\}_{i,j=1}^{N,N_i}$ and the weights $\{\omega_{ij}\}_{i,j=1}^{N,N_i}$ as $h_t^s(\bar{x}_{ij}) = \arg\max_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(\bar{x}_{ij}) w_{ij}$, where $\bar{x}_{ij} = \{\hat{y}_{ik} | x_{ik} \in \text{Nbr}(x_{ij})\}$ are the predicted labels of the neighbors $\text{Nbr}(x_{ij})$ of the current instance x_{ij}.

The classifier which has lower training error is selected as the weak *instance* classifier $h_t(x_{ij})$,

$$h_t(x_{ij}) = \begin{cases} h_t^d(x_{ij}) \text{ if } & \epsilon^d < \epsilon^s \\ h_t^s(\bar{x}_{ij}) & \text{otherwise} \end{cases}, \tag{10}$$

where $\epsilon^d = \sum_{i,j} \omega_{ij} |h_t^d(x_{ij}) - l_i|$ and $\epsilon^s = \sum_{i,j} \omega_{ij} |h_t^s(\bar{x}_{ij}) - l_i|$ are the training errors of the weak *data* classifier $h_t^d(x_{ij})$ and the weak *spatial* classifier $h_t^s(\bar{x}_{ij})$, respectively. This is the major difference of the proposed Spatial-MILBoosting algorithm and traditional MILBoost algorithm [19, 35].

The parameter λ_t is determined using a line search as $\lambda_t = \arg\max_\lambda L(H + \lambda h_t)$. Then, the instance classifier $H(\cdot)$ is updated by $H(\cdot) \leftarrow H(\cdot) + \lambda_t h_t(\cdot)$.

3.2 Appearance Model

Since the appearance of the object instances (also the backgrounds) are similar within each video V^n while exhibiting large variations across $\mathcal{V}$, we independently learn the color distributions of the target objects and the backgrounds for each video V^n.

In detail, with a segmentation $\mathcal{B}$ for $\mathcal{V}$, we estimate two color Gaussian Mixture Models (GMMs) for the target objects and the backgrounds of each video V^n, denoted as $\mathbf{h}_1^n$ and $\mathbf{h}_0^n$, respectively. The corresponding data term based on the appearance model in Eq(1) is defined as

$$D_j^2(b_{ij}^n) = -\log \mathbf{h}_{b_{ij}^n}^n(s_{ij}^n), \tag{11}$$

where $D_j^2(b_{ij}^n)$ measures the contribution of labeling superpixel s_{ij}^n to be b_{ij}^n, based on the appearance model learned from video V^n.

3.3 Consistency Term

The consistency term is composed of an intra-frame consistency model and an inter-frame consistency model, and is leveraged to constrain the segmentation labels to be both spatially and temporally consistent.

Intra-frame Consistency Model. The intra-frame consistency model encourages the spatially adjacent superpixels in the same frame to have the same label. In Eq(1), the consistency term computed between spatially adjacent superpixels s_{ij}^n and s_{ik}^n in frame f_i^n of video V^n is defined as

$$S_{jk}^1(b_{ij}^n, b_{ik}^n) = \delta(b_{ij}^n, b_{ik}^n) \exp\left(-||\mathbf{I}_{ij}^n - \mathbf{I}_{ik}^n||_2^2\right), \tag{12}$$

where $\mathbf{I}$ is the color vector of the superpixel, and b_{ij}^n and b_{ik}^n are the segmentation labels of s_{ij}^n and s_{ik}^n. $\delta(\cdot)$ denotes the Dirac delta function, which is 0 when $b_{ij}^n = b_{ik}^n$, and 1 otherwise.

Inter-frame Consistency Model. The inter-frame consistency model encourages the temporally adjacent superpixels in consecutive frames to have the same label. In Eq(1), the consistency term computed between temporally adjacent superpixels s_{ij}^n and s_{uk}^n in consecutive frames of video V^n is defined as

$$S_{jk}^2(b_{ij}^n, b_{uk}^n) = \delta(b_{ij}^n, b_{uk}^n) \exp\left(-||\mathbf{c}_{ij}^n - \mathbf{c}_{uk}^n||_1\right), \tag{13}$$

where $\mathbf{c}$ is the context vector of the superpixel, and b_{ij}^n and b_{uk}^n are the segmentation labels of s_{ij}^n and s_{uk}^n. s_{uk}^n is the temporal neighbor of s_{ij}^n, *i.e.*, its corresponding next superpixel $\overrightarrow{s}_{ij}^n$ in frame f_{i+1}^n or previous superpixel $\overleftarrow{s}_{ij}^n$ in frame f_{i-1}^n.

4 Optimization

The proposed approach is bootstrapped from a few manually annotated relevant and irrelevant frames (*e.g.*, usually 1 to 3), and an objectness measure [2] and a saliency measure [14] to initialize the segmentation $\mathcal{B}$ and the discriminative probability map set $\mathcal{P}$ of $\mathcal{V}$. We proceed to start the first round learning of the spatiotemporal auto-context model, and propagate the superpixel labels estimated from the learned auto-context classier $H(\cdot)$ to frame-level labels $\mathcal{L}$ of $\mathcal{V}$ through the Spatial-MILBoosting algorithm. We then update the spatiotemporal auto-context model together with the appearance model and consistency term, and perform energy minimization on Eq(1) by using min-cut [6, 7] to obtain an updated segmentation $\mathcal{B}$ of $\mathcal{V}$.

The learning of the spatiotemporal auto-context model (the object discovery), and the minimization of the energy function in Eq(1) (the object co-segmentation) are iteratively performed until convergence, which returns not only a frame-level label $\mathcal{L}$ of $\mathcal{V}$ and a segmentation $\mathcal{B}$ of $\mathcal{V}$, but also a spatiotemporal auto-context model.

Object Discovery. The object discovery is to identify the relevant frames containing the target objects from multiple videos $\mathcal{V}$. As we obtained a current frame-level labels $\mathcal{L}$, segmentation $\mathcal{B}$, and discriminative probability map set $\mathcal{P}$ estimated by the spatiotemporal auto-context model from the previous iteration, the probability of frame f_i^n containing the target object is updated as

$$p_i^n = 1 - (1 - p_i^n(1))(1 - p_i^n(0)), \tag{14}$$

where $p_i^n(1)$ and $p_i^n(0)$ are the probabilities of the object and background of f_i^n being positive, respectively. They are calculated by Eq(4) and Eq(6) above in Sec. 3.1. Then, the label l_i^n indicating if f_i^n is relevant can be predicted by binarizing p_i^n. l_i^n equals to 1 when f_i^n is relevant, and 0 irrelevant. In this way, the label l_i^n can be inferred from the probabilities of the object and background inside f_i^n indicating if they contain the target objects; while the probability of the object (or background) can be inferred from the probabilities of the superpixels inside it denoting if they belong to the target object.

Object Co-segmentation. The video object co-segmentation is to simultaneously find a superpixel-level labeling $\mathcal{B}$ for the relevant frames identified from $\mathcal{V}$. As we obtain a current frame-level labels $\mathcal{L}$, segmentation $\mathcal{B}$ and discriminative probability map set $\mathcal{P}$ estimated by the spatiotemporal auto-context model, we can update the video independent spatiotemporal auto-context model. Naturally, the spatiotemporal contextual information across multiple videos $\mathcal{V}$ are leveraged for the segmentation of each frame. The new segmentation B^n of each video V^n also serves to update the corresponding video dependent appearance model and consistency term. We then minimize the energy function in Eq(1) using min-cut [6, 7] to obtain the new segmentation $\mathcal{B}$ of $\mathcal{V}$.

5 Experiments and Discussions

We conduct extensive experiments to evaluate our method on three datasets, including the SegTrack dataset [30], the video co-segmentation dataset [12, 27, 29], and a new 10-categories video object co-segmentation dataset collected by ourselves.

5.1 Evaluation on the SegTrack v1 and v2 datasets

The SegTrack (v1 [30] and v2 [17]) is a video segmentation dataset consisting of 8 videos containing one object and 6 videos containing multiple adjacent/interacting objects, with full pixel-level annotations on the objects at each frame. As our method focuses on single object segmentation, we test our method on the 8 videos containing one object. By initializing all frames as relevant, we segment each video using our method.

We first compute the average per-frame pixel error rate for each video, and compare it with 8 other methods [8, 12, 17, 20, 22, 23, 41] on 3 videos from SegTrack v1 dataset [30], as summarized in Table 2. We also compare the average intersection-over-union score of our method with 4 video segmentation methods [12, 16, 17] on the videos from SegTrack v2 dataset [17], as summarized in Table 3.

Table 2. The per-frame pixel error rates of our method and 8 other methods [8,12,17,20,22,23,41] on SegTrack v1 dataset [30]. Lower values are better.

Video	Ours	[23]	[17]-1	[17]-2	[41]	[20]	[22]	[12]	[8]
girl	**1053**	3859	1573	1564	1488	1698	5683	5777	7595
birdfall	**152**	217	188	242	155	189	468	305	468
parachute	**189**	855	339	328	220	221	1595	1202	1113

Table 3. The intersection-over-union scores of our method and 4 other video segmentation methods [12, 16, 17] on SegTrack v2 dataset [17]. Higher values are better.

Algorithm	girl	birdfall	parachute	frog	worm	soldier	monkey	bird of paradise
Ours	**90.5**	**70.3**	92.4	**83.1**	80.4	**85.3**	**89.8**	**94.5**
[17]-1	89.1	62.0	93.2	65.8	75.6	83.0	84.1	88.2
[17]-2	89.2	62.5	93.4	72.3	82.8	83.8	84.8	94.0
[16]	87.7	49.0	**96.3**	0	**84.4**	66.6	79.0	92.2
[12]	31.9	57.4	69.1	67.1	34.7	66.5	61.9	86.8

The per-frame pixel error rate is the number of pixels misclassified according to the ground truth segmentation, and is calculated as $error = N_{seg \oplus gt}$. The intersection-over-union is calculated as $N_{seg \cap gt}/N_{seg \cup gt}$, where $N_{seg \cap gt}$ and $N_{seg \cup gt}$ are the pixel numbers of the intersection and the union of the segmentation result and the ground truth segmentation, respectively. The [17]-1 and [17]-2 in Table 2 and Table 3 denote the original method [17], and the method [17] plus a refinement process using composite statistical inference, respectively. Some qualitative example results of our method are presented in Fig. 5 of the supplementary material.

As the results in Table 2 shown, our method outperforms the other 8 methods on the 3 videos. The results in Table 3 showed that our method is superior among the 4 other methods on 6 videos, but underperforms the other methods on 2 videos. The intersection-over-union score on parachute is slightly lower because of the complex background caused by difficult lighting conditions. The worm is difficult to segment since the boundaries between the worms and the background in some frames are too weak. For the birdfall, the frames are complex due to the cluttered background and the small size of the birds. In general, as the results shown, our method has the ability to segment the objects with certain variations in appearance (bird of paradise), shape (girl and frog), size (soldier), and backgrounds (parachute), but has encountered some difficulties when the objects are too small (birdfall), or the boundaries between the objects and the background are too weak (worm).

5.2 Evaluation on the video co-segmentation dataset

We also test our method on videos of 3 categories, *i.e.*, 4 videos of the Cha-cha-cha category from Chroma dataset [29], 3 videos of the kite surfing category and 3 videos of the ice dancing category both from [12] and [27]. Because all frames from all videos of each category contain the target objects, we treat all frames of each category as relevant, and simultaneously segment the videos of each category using our method.

Table 4. The labeling accuracy of our method and two video object co-segmentation methods [13, 27] on 4 videos of the Cha-cha-cha category. Higher values are better.

Algorithm	cha.1	cha.2	cha.3	cha.4
Ours	**97.1**	**96.9**	**97.0**	**97.5**
[13] + [21]	96	96	95	96
[27]	61	81	56	74

Table 5. Labeling accuracy on the kite surfing and ice dancing categories.

Algorithm	kite.1	kite.2	kite.3	ice.1	ice.2	ice.3
Ours	93.7	94.1	95.8	97.2	96.5	98.1

We compute the average labeling accuracy on each video of the Cha-cha-cha category, and compare them with 2 other video object co-segmentation methods [13, 27], as presented in Table 4. Since the method presented in [13] produces the results in terms of dense trajectories, they use the method in [21] to turn their trajectory labels into pixel labels for comparison.

The labeling accuracy is calculated as $N_{seg \odot gt}/N_{total}$, *i.e.*, it is the ratio of the number of pixels classified correctly in accordance with the ground truth segmentation to the total number of pixels. We also present some qualitative results of our method compared with [13, 27] on the Cha-cha-cha category in Fig. 3 (a). These results showed that our method outperforms the other 2 video object co-segmentation methods [13,27], and is not limited to the initial segmentation generated by combing the objectness and saliency measures that the method in [27] is sensitive to.

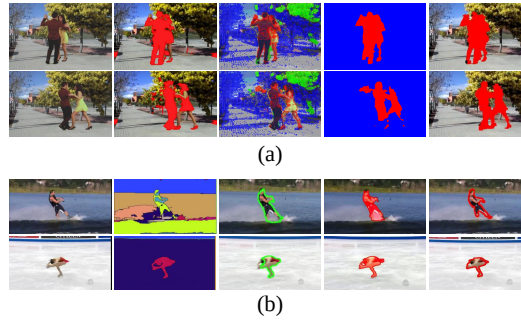

Fig. 3. Some qualitative results of our method compared with other methods [12, 13, 16, 27, 27]. (a) From left to right: original frames, results of [27], [13], [13] plus [21], and our results on the Cha-cha-cha category. (b) From left to right: original frames, results of [12], [16], [27], and our results on the kite surfing and ice dancing categories.

The average labeling accuracies computed on videos of the kite surfing and ice dancing categories by our method are presented in Table 5. We also present some qualitative results of our method compared with [12, 16, 27] on the two categories in Fig. 3 (b). They showed that our method compares favorably or is on par with [12, 16, 27].

5.3 Evaluation on the new video object co-segmentation dataset

New video object co-segmentation dataset. To exactly evaluate the efficacy of our method and to establish a benchmark for future research, we have collected 10 categories of 101 publicly available Internet videos, in which some videos include irrelevant frames. We manually assign each frame a label (1 for relevant and 0 for irrelevant), and also manually assign pixel-wise ground truth foreground labels for each relevant frame. The statistical details of the new dataset are given in Table 6. We present some

Table 6. The new video co-segmentation dataset. "Video (R./I.)" denotes the numbers of all videos, videos only containing the relevant frames, and videos containing irrelevant frames; "Frame (R./I.)" denotes the numbers of all frames, relevant frames, and irrelevant frames in videos of each category.

Category	Video (R./I.)	Frame (R./I.)	**Category**	Video (R./I.)	Frame (R./I.)
airplane	11(4/7)	1763(1702/61)	balloon	10(4/6)	1459(1394/65)
bear	11(6/5)	1338(1282/56)	cat	4(3/1)	592(578/14)
eagle	13(12/1)	1703(1665/38)	ferrari	12(9/3)	1272(1244/28)
figure skating	10(7/3)	1173(1115/58)	horse	10(5/5)	1189(1134/55)
parachute	10(4/6)	1461(1421/40)	single diving	10(0/10)	1448(1372/76)

Table 7. The discovery performance of our method by varying the number of manually annotated frames (the number in the 1st row). The number in the table is the misclassified frames when 1, 2, and 3 labeled frames are provided.

Category	1	2	3	**Category**	1	2	3	**Category**	1	2	3
airplane	20	10	0	balloon	13	4	3	bear	3	3	2
cat	4	5	5	eagle	23	12	8	ferrari	11	7	6
figure skating	0	0	0	horse	5	1	1	parachute	14	10	2
single diving	18	13	5	-	-	-	-	-	-	-	-

example relevant and irrelevant frames for each category of the new dataset in Fig. 6 of the supplementary material. The objects in videos of each category are of the common category, but exhibit large differences in appearance, size, shape, viewpoint, and pose.

Performance evaluation. To better understand the contributions of the different aspects of our proposed method, we perform an ablative study. To this end, in addition to the proposed method (denoted V-1), we implemented a variant where Spatial-MILBoosting was replaced by MILBoost [35] (denoted V-2).

We first evaluate the discovery performance of our method by varying the number of manually annotated relevant and irrelevant frames. In our experiments, the number of manually annotated relevant and irrelevant frames of each video are set from 1 to 3, and they are randomly selected from each video given the ground truth frame-level labels. We present the number of misclassified frames of each category tested on the new video co-segmentation dataset in Table 7. As the results shown, our method works well when just provide each video 1 relevant or irrelevant frame, and can identify almost all the relevant frames from multiple videos when we provide 3 relevant and irrelevant frames. This validated the efficacy of the spatiotemporal auto-context model learned through the Spatial-MILBoosting algorithm.

Table 8 presents the average intersection-over-union scores of two versions of our method tested on each category of the new dataset. Some qualitative results of two versions of our method on videos of each category are presented in Fig. 4. They demonstrate the advantages of our method. In addition, it also demonstrates the advantages of the Spatial-MILBoosting algorithm, which considers the spatial relationship of neighboring superpixels while predicting the segmentation label of superpixel.

Table 8. Ablative study comparing Spatial-MILBoosting vs. MILBoost [35] on intersection-over-union on the new video co-segmentation dataset.

Category	V-1	V-2	Category	V-1	V-2	Category	V-1	V-2
airplane	86.4	84.7	balloon	94.6	93.9	bear	90.5	89.3
cat	92.1	89.4	eagle	89.5	86.2	ferrari	87.7	86.3
figure skating	88.5	86.9	horse	92.0	90.7	parachute	94.0	91.7
single diving	87.7	85.2	-	-	-	-	-	-

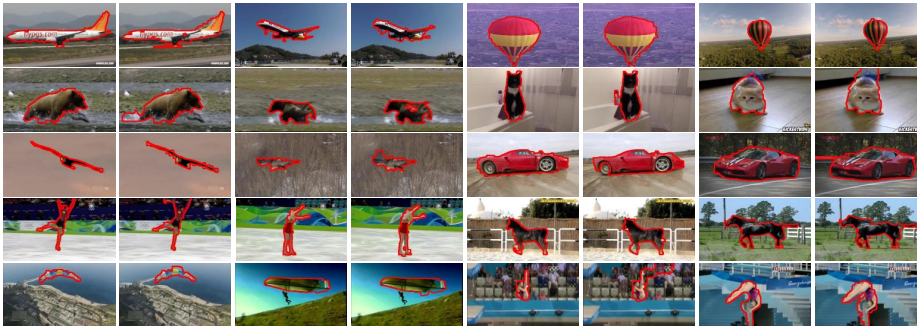

Fig. 4. Qualitative results of two versions of our method tested on each category of the new dataset. The 1, 3, 5 and 7 columns: results of V-1; the 2, 4, 6 and 8 columns: results of V-2.

To summarize, as shown above, our method has the capability of discovering the relevant frames from multiple videos containing irrelevant frames, and clearly co-segmenting the common objects from them.

6 Conclusion

We presented a spatiotemporal energy minimization formulation to simultaneously discover and co-segment a common category of objects from multiple videos containing irrelevant frames, which only requires extremely weak supervision (*i.e.*, 1 to 3 frame-level labels). Our formulation incorporates a spatiotemporal auto-context model to capture the spatiotemporal contextual information across multiple videos. It facilitates both the object discovery and co-segmentation through a MIL algorithm with spatial reasoning. Our method overcomes an important limitation of previous video object co-segmentation methods, which assume all frames from all videos contain the target objects. Experiments on three datasets demonstrated the superior performance of our proposed method.

Acknowledgements. This work was partly supported by China 973 Program Grant 2012CB316400, and NSFC Grant 61228303. Le Wang was supported by the Ph.D. Short-term Academic Visiting Program of Xi'an Jiaotong University. Dr. Gang Hua was partly supported by US National Science Foundation Grant IIS 1350763, a Google Research Faculty Award, and GHs start-up funds from Stevens Institute of Technology.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. TPAMI 34(11), 2274–2282 (2012)
2. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR. pp. 73–80 (2010)
3. Avidan, S.: SpatialBoost: Adding spatial reasoning to adaboost. In: ECCV. pp. 386–396 (2006)
4. Bai, X., Wang, J., Simons, D., Sapiro, G.: Video SnapCut: robust video object cutout using localized classifiers. In: ACM Trans. on Graphics. vol. 28, p. 70 (2009)
5. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: iCoseg: Interactive co-segmentation with intelligent scribble guidance. In: CVPR. pp. 3169–3176 (2010)
6. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient ND image segmentation. IJCV 70(2), 109–131 (2006)
7. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. TPAMI 26(9), 1124–1137 (2004)
8. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV. pp. 282–295 (2010)
9. Chen, D.J., Chen, H.T., Chang, L.W.: Video object cosegmentation. In: ACM Multimedia. pp. 805–808 (2012)
10. Chiu, W.C., Fritz, M.: Multi-class video co-segmentation with a generative multi-video model. In: CVPR. pp. 321–328 (2013)
11. Dai, J., Wu, Y.N., Zhou, J., Zhu, S.C.: Cosegmentation and cosketch by unsupervised learning. In: ICCV (2013)
12. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: CVPR. pp. 2141–2148 (2010)
13. Guo, J., Li, Z., Cheong, L.F., Zhou, S.Z.: Video co-segmentation for meaningful action extraction. In: ICCV (2013)
14. Harel, J., Koch, C., Perona, P., et al.: Graph-based visual saliency. NIPS pp. 545–552 (2006)
15. Joulin, A., Bach, F., Ponce, J.: Discriminative clustering for image co-segmentation. In: CVPR. pp. 1943–1950 (2010)
16. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV. pp. 1995–2002 (2011)
17. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV (2013)
18. Liu, D., Chen, T.: A topic-motion model for unsupervised video object discovery. In: CVPR. pp. 1–8 (2007)
19. Liu, D., Hua, G., Chen, T.: A hierarchical visual model for video object summarization. TPAMI 32(12), 2178–2190 (2010)
20. Ma, T., Latecki, L.J.: Maximum weight cliques with mutex constraints for video object segmentation. In: CVPR. pp. 670–677 (2012)
21. Ochs, P., Brox, T.: Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In: ICCV. pp. 1583–1590 (2011)
22. Ochs, P., Brox, T.: Higher order motion models and spectral clustering. In: CVPR. pp. 614–621 (2012)
23. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV (2013)
24. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR. pp. 3282–3289 (2012)
25. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: CVPR. pp. 1939–1946 (2013)

26. Rubinstein, M., Liu, C., Freeman, W.T.: Annotation propagation in large image databases via dense image correspondence. In: ECCV, pp. 85–99 (2012)
27. Rubio, J.C., Serrat, J., López, A.: Video co-segmentation. In: ACCV. pp. 13–24 (2012)
28. Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: CVPR. pp. 2483–2490 (2013)
29. Tiburzi, F., Escudero, M., Bescós, J., Martínez, J.M.: A ground truth for motion-based video-object segmentation. In: ICIP. pp. 17–20 (2008)
30. Tsai, D., Flagg, M., Rehg, J.: Motion coherent tracking with multi-label MRF optimization. BMVC (2010)
31. Tu, Z.: Auto-context and its application to high-level vision tasks. In: CVPR. pp. 1–8 (2008)
32. Tuytelaars, T., Lampert, C.H., Blaschko, M.B., Buntine, W.: Unsupervised object discovery: A comparison. IJCV 88(2), 284–302 (2010)
33. Vicente, S., Kolmogorov, V., Rother, C.: Cosegmentation revisited: Models and optimization. In: ECCV. pp. 465–479 (2010)
34. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: CVPR. pp. 2217–2224 (2011)
35. Viola, P., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: NIPS. pp. 1417–1424 (2005)
36. Wang, L., Xue, J., Zheng, N., Hua, G.: Automatic salient object extraction with contextual cue. In: ICCV. pp. 105–112 (2011)
37. Wang, L., Xue, J., Zheng, N., Hua, G.: Concurrent segmentation of categorized objects from an image collection. In: ICPR. pp. 3309–3312 (2012)
38. Wang, L., Hua, G., Xue, J., Gao, Z., Zheng, N.: Joint segmentation and recognition of categorized objects from noisy web image collection. TIP (2014)
39. Xu, L., Jia, J., Matsushita, Y.: Motion detail preserving optical flow estimation. TPAMI 34(9), 1744–1757 (2012)
40. Xue, J., Wang, L., Zheng, N., Hua, G.: Automatic salient object extraction with contextual cue and its applications to recognition and alpha matting. PR 46(11), 2874–2889 (2013)
41. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: CVPR. pp. 628–635 (2013)
42. Zhao, G., Yuan, J., Hua, G.: Topical video object discovery from key frames by modeling word co-occurrence prior. In: CVPR. pp. 1602–1609 (2013)