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 · IAIR Institute of Artificial Intelligence and Robotics · STEVENS INSTITUTE of TECHNOLOGY THE INNOVATION UNIVERSITY · Google

1. Motivations & Contributions

Motivation

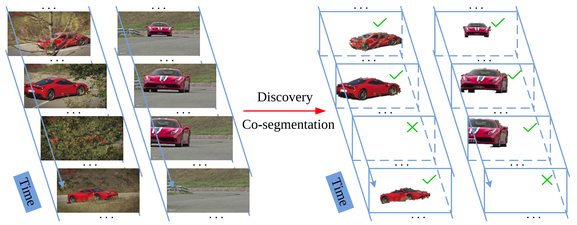
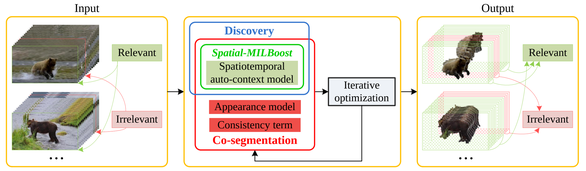


Fig. Problem setting: **Input** - multiple videos capture a common category of objects. Some of which may contain irrelevant frames. **Output** - a label for each frame indicating if it is relevant, and a detailed pixel labeling of the common object for each relevant frame identified.

Contributions

– We propose a method to address the problem of simultaneous discovery and co-segmentation of a common category of objects from multiple videos containing irrelevant frames.
– We model the spatiotemporal contextual information across multiple videos by a spatiotemporal auto-context model learned from a Spatial-MILBoosting algorithm.
– We collect a new 10-categories video co-segmentation dataset with ground truth frame-level labels for all frames and pixel-wise segmentation labels for all relevant frames.

2-1. Framework



2-2. Principal Notation

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

2-3. Problem Formulation

Spatiotemporal auto-context model — Data term — Appearance model

$$\mathrm{E}(\mathcal{B}) = \sum_{s_{ij}^n \in \mathcal{V}} D_j^1(b_{ij}^n) + \sum_{s_{ij}^n \in V^n} D_j^2(b_{ij}^n) + \sum_{s_{ij}^n, s_{ik}^n \in \mathcal{N}_j} S_{jk}^1(b_{ij}^n, b_{ik}^n) + \sum_{s_{ij}^n, s_{uk}^n \in \bar{\mathcal{N}}_j} S_{jk}^2(b_{ij}^n, b_{uk}^n), \quad (1)$$

Intra-frame consistency model — Consistency term — Inter-frame consistency model

$$n = 1,\ldots,N, i = 1,\ldots,N^n, j = 1,\ldots,N_i^n$$

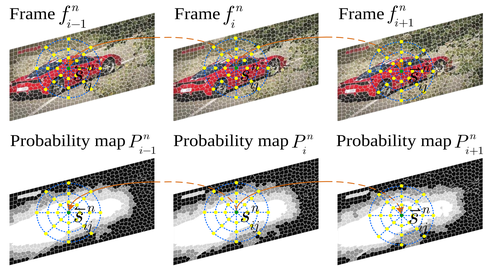
Data term

- measuring the cost of labeling s_{ij}^n to be b_{ij}^n from a spatio-temporal auto-context model and an appearance model, respectively.

Consistency term

- constraining the segmentation labels to be both spatially and temporally consistent from an intra-frame consistency model and an inter-frame consistency model.

3-1. Spatiotemporal Auto-context model



- Context feature of s_{ij}^n : $\mathbf{c}_{ij}^n = \{\{\tilde{p}_{ij}^n(\mathrm{k})\},\{p_{ij}^n(\mathrm{k})\},\{\bar{p}_{ij}^n(\mathrm{k})\}\}_{k=1}^{N_c}$
- $\mathbf{S} = \{\{\mathbf{C}_i^n(\alpha), l_i^n(\alpha)\} \,|\, n = 1,\ldots,N; i = 1,\ldots,N^n; \alpha = 0,1\}$, where $\mathbf{C}_i^n = \{\mathbf{c}_{ij}^n\}_{j=1}^{N_i^n}$ are the context features in f_i^n, $\mathbf{C}_i^n(\alpha)$ are the context features in the object or background of f_i^n.
- We treat $\mathbf{C}_i^n(\alpha)$ as a bag, and $\mathbf{c}_{ij}^n$ as an instance. $l_i^n(\alpha)$ is the label of $\mathbf{C}_i^n(\alpha)$. The classifier $H(\cdot)$ is learned on $\mathbf{S}$ using Spatial-MILBoosting.
- The new probability of s_{ij}^n being positive is given as

$$p_{ij}^n = \frac{1}{1+\exp(-H(\mathbf{c}_{ij}^n))}. \quad (2)$$

- The data term based on the spatiotemporal auto-context model is defined as $D_j^1(b_{ij}^n) = -\log p_{ij}^n$. (3)
- The probability of the object or background (bag) of frame f_i^n being positive is a "Noisy OR" defined as

$$p_i^n(\alpha) = 1 - \prod_{j=1}^{N_i^n(\alpha)} (1 - p_{ij}^n). \quad (4)$$

3-2. Appearance Model

- Estimate two color GMMs for the target objects and the background of each video V^n, denoted as $\mathbf{h}_1^n$ and $\mathbf{h}_0^n$.
- The corresponding data term based on the appearance model is defined as

$$D_j^2(b_{ij}^n) = -\log \mathbf{h}_{b_{ij}^n}^n(s_{ij}^n). \quad (5)$$

3-3. Consistency Term

Intra - frame consistency model

- The consistency between s_{ij}^n and s_{ik}^n in f_i^n of V^n

$$S_{jk}^1(b_{ij}^n, b_{ik}^n) = \delta(b_{ij}^n, b_{ik}^n)\exp(-\| \mathbf{I}_{ij}^n - \mathbf{I}_{ik}^n \|_2^2). \quad (6)$$

Inter - frame consistency model

- The consistency between temporally adjacent superpixels s_{ij}^n and s_{uk}^n in consecutive frames of V^n

$$S_{jk}^2(b_{ij}^n, b_{uk}^n) = \delta(b_{ij}^n, b_{uk}^n)\exp(-\| \mathbf{c}_{ij}^n - \mathbf{c}_{uk}^n \|_1). \quad (7)$$

4. Iterative Optimization

Object Discovery

- The probability of f_i^n containing the object is

$$p_i^n = 1 - (1 - p_i^n(1))(1 - p_i^n(0)). \quad (8)$$

Object Co - segmentation

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

5. Spatial-MILBoosting Algorithm

Input : training set $\{\mathbf{x}_i, l_i\}_{i=1}^N$, where $\mathbf{x}_i = \{x_{ij}\}_{i=1}^{N_i}$, $l_i \in \{0,1\}$

1. Initialize weights $w_{ij} = 2*(l_i - 0.5)$, and 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 $\bar{x}_{ij} = \{\hat{y}_{ik} \,|\, x_{ik} \in \mathrm{Nbr}(x_{ij})\}$
   b. Train weak data classifier
   $h_t^d(x_{ij}) = \mathrm{argmax}_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(x_{ij}) w_{ij}$
   c. Train weak spatial classier
   $h_t^s(\bar{x}_{ij}) = \mathrm{argmax}_{\hat{h}(\cdot)} \sum_{i,j} \hat{h}(\bar{x}_{ij}) w_{ij}$
   d. Set $\varepsilon^d = \sum_{i,j} w_{ij} |h_t^d(x_{ij}) - l_i|$ and $\varepsilon^s = \sum_{i,j} w_{ij} |h_t^s(\bar{x}_{ij}) - l_i|$
   e. Set $h_t(x_{ij}) = \begin{cases} h_t^d(x_{ij}) & \text{if } \varepsilon^d < \varepsilon^s \\ h_t^s(\bar{x}_{ij}) & \text{otherwise} \end{cases}$
   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 $\hat{y}_{ij} = H(x_{ij}) = \hat{y}_{ij} + \lambda_t h_t(x_{ij})$
   h. Compute the instance probability
   $q_{ij} = 1 / (1 + \exp(-\hat{y}_{ij}))$
   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 $H(x_{ij}) = \sum_{t=1}^T \lambda_t h_t(x_{ij})$

6. Experiments

A. Object Segmentation from a Single Video

Table 1. The per-frame pixel error rates of ours and 8 other methods on SegTrack.

Algorithm	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 2. The intersection-over-union scores of ours and 4 other methods on SegTrack.

Algorithm	girl	birdfall	parachute	frog	worm	soldier	monkey	bird
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

B. Object Co-segmentation from Multiple Videos

Table 3. The labeling accuracy of our method and two video object co-segmentation methods [13, 27] on 4 videos of the Cha-cha-cha category.

Algorithm	Cha-cha-cha1	Cha-cha-cha2	Cha-cha-cha3	Cha-cha-cha4
Ours	97.1	96.9	97.0	97.5
[13] + [21]	96	96	95	96
[27]	61	81	56	74

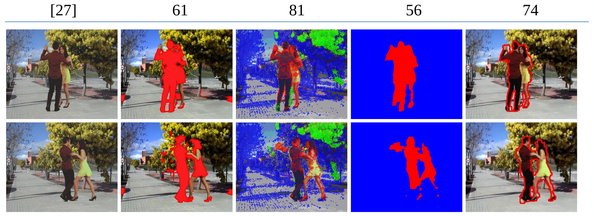


Fig. Comparison results on the Cha-cha-cha category. From left to right: original frames, results of [27], [13], [13] plus [21], and ours.

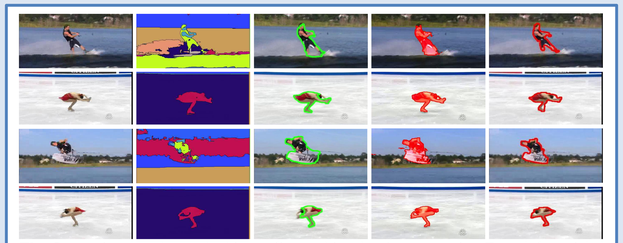


Fig. Comparison results on the kite surfing and ice dancing categories. From left to right: original frames, results of [12], [16], [27], and ours.

C. New Video Co-segmentation Dataset

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

Fig. The 1st, 3rd, 6th and 8th columns are example relevant frames; the 2rd, 4th, 7th and 9th columns are the pixel-wise object segmentation labels; the 5th and 10th columns are example relevant frames.

D. Object Discovery and Co-segmentation

Table 6. The discovery performance of our method by varying the number of manually annotated frames (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
airplane	20	10	0	ferrari	11	7	6
balloon	13	4	3	figure skating	0	0	0
bear	3	3	2	horse	5	1	1
cat	4	5	5	parachute	14	10	2
eagle	23	12	8	single diving	18	13	5

Table 7. 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
airplane	86.4	84.7	ferrari	87.7	86.3
balloon	94.6	93.9	figure skating	88.5	86.9
bear	90.5	89.3	horse	92.0	90.7
cat	92.1	89.4	parachute	94.0	91.7
eagle	89.5	86.9	single diving	87.7	85.2

V-1 – the proposed method.
V-2 – a variant where Spatial-MILBoosting was replaced by MILBoost [35].

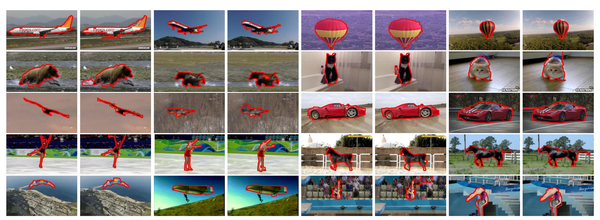


Fig. Qualitative results of two versions of our method tested on each category of the new dataset. The 1, 3, 5 and 7 columns: results of V-1; the 2, 4, 6 and 8 columns: results of V-2.