TRANSDUCTIVE VIDEO CO-SEGMENTATION ON THE TEMPORAL TREES

Zhihui Fu, Botao Wang, Student Member, IEEE, *Hongkai Xiong,* Senior Member, IEEE

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

ABSTRACT

This paper proposes a novel multi-component video co-segmentation approach to simultaneously separate the foreground from the background in the video frames. To capture the variance of appearance of the foreground object, a multi-component foreground model is developed. Each component of the model characterizes a specific viewpoint/pose/appearance of the foreground object. To learn the parameters of the multi-component model, a transductive learning algorithm is leveraged to "transfer" the information of the labeled frames to the unlabeled frames in a tree-structured model, namely, *temporal tree*. Each branch of the temporal tree consists of the exemplars of a foreground component, and a transductive support vector regressor is capable of being trained. Experiments show that the proposed method outperforms quite a few state-of-the-art video segmentation algorithms in public benchmark.

Index Terms— Video segmentation, co-segmentation, transductive learning, mixture model.

1. INTRODUCTION

Video segmentation aims at separating video frames into consistent regions with both spatial and temporal consistence. It serves as an imperative phase for various computer vision tasks, e.g., object detection [1], visual tracking [2] , and action recognition [3].

Current video segmentation algorithms can be broadly divided into two categories: *unsupervised approaches* and *supervised approaches*. Unsupervised approaches [4, 5] group the pixels into spatially and temporally coherent clusters based on the visual and motion cues of the video. However, they often result in over-segmentation for complex videos because of the ambiguity of visual features. Up to now, the popular motion segmentation [6–8] mainly focuses on the objects which exhibit different motion patterns in comparison with the background. Obviously, it might fail if the objects of interest are static.

Supervised segmentation [9, 10], also known as interactive segmentation, is devoted to segmenting the objects which could be hinted in a few frames. The task is more well-defined – to segment the foreground objects of interest from the background in the frames. It stimulated the notion of image co-segmentation [11, 12], namely, automatically segmenting common foreground in multiple images. A couple of methods [13, 14] borrow the idea of image co-segmentation for video segmentation, however, would require multiple videos which contain similar foreground and diverse background. It is worth mentioning that this paper is dedicated to co-segmenting the foreground of the frames in a single video clip. As a matter of fact, the main challenge of video co-segmentation is to separate the foreground from the background, which are both high-correlated across frames. Within the supervised video segmentation, a small number of annotated frames could provide critical prior to disambiguate them.

The contributions of this paper are two-fold. First, a multi-component video co-segmentation approach is proposed. On the one hand, it simultaneously segments the foreground regions in the frames of a video clip by maximizing both the inter-frame similarity between the foregrounds and the intra-frame dissimilarity between the foreground and the background. On the other hand, it enforces temporal consistency regularization to video co-segmentation, so as to keep the region both visually consistent and temporally coherent. Unlike [14, 15], it uses multiple foreground models to capture the variances of the foreground object which exhibits changes in pose and appearance.

Second, the parameters of the multiple foreground models are learned with a tree structure by a transductive learning algorithm. Each path in the temporal tree defines a component model, which consists of visually similar and temporally adjacent frames. It is constructed for all the frames to predict the foreground masks of the unlabeled frames by transferring the prior of the labeled frames to the unlabeled frames. It is transferred by fitting a support vector regressor to maximize the margin of the prediction error of the foreground region. The experimental results show that the proposed method outperforms state-of-the-art video segmentation and image co-segmentation algorithms in public benchmarks.

The rest of this paper is organized as follows: Section 2 presents the proposed video co-segmentation framework. Section 3 illustrates the transductive learning algorithm to fit the hyperparameters of the foreground model. Section 4 provides the experiment results. Section 5 concludes the paper.

The work was supported in part by the NSFC, under grants 61425011, U1201255, and 61271218.

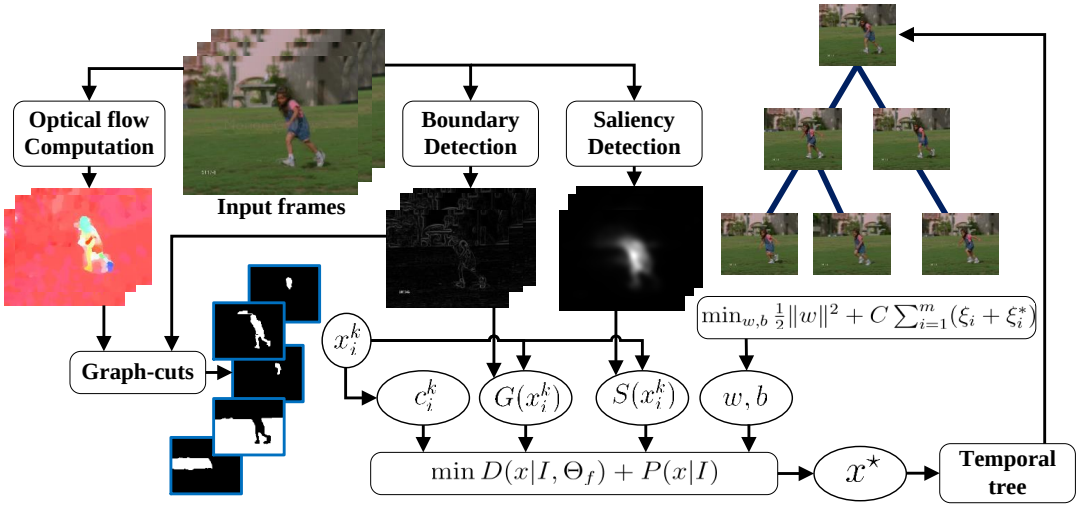


Fig. 1. The proposed video co-segmentation framework.

2. TEMPORAL COHERENT VIDEO CO-SEGMENTATION

2.1. System Overview

The proposed framework is illustrated in Fig. 1. The video sequence, denoted as $\{I_t\}_{t=1}^N$, consists of labeled frames and unlabeled frames, where N is the total number of frames. Without loss of generalization, let us assume that the first $L : 1 \leq L \ll N$ frames $\{I_t\}_{t=1}^L$ are labeled. Hence, the task is to predict the foreground masks of the unlabeled frames $\{I_t\}_{t=L+1}^N$. A *segmentation* of a frame is defined with a binary matrix $x \in \{0,1\}^{H\times W}$, where H and W is the height and the width of the frames, respectively. $x(u,v) = 1$ if the pixel at (u,v) is the foreground, and $x(u,v) = 0$ otherwise.

The quality of a segmentation is measured by an energy function parameterized by the foreground hyperparameter Θ_f. To obtain the hyperparameters, which characterize both the visual appearance and the temporal coherence between the foreground region and the background region, a transductive learning procedure is performed to iteratively optimize the hyperparameters via updating the temporal tree.

2.2. Foreground Energy Function

The foreground energy of a frame is defined as

$$E(x|I,\Theta_f) = D(x|I,\Theta_f) + \alpha P(x|I), \tag{1}$$

which is composed of two terms: the *divergence term* $D(x|I,\Theta_f)$ and the *prior term* $P(x|I)$. The divergence term measures the visual similarity of the foreground x with the foreground model, which is parameterized by Θ_f. The prior term $P(x|I)$ evaluates the quality of the segmentation based on the low-level features of the image.

The divergence term $D(x|I,\Theta_f)$ is defined as

$$D(x|I,\Theta_f) = f(c(x),\Theta_f) - \beta f(c(\bar{x}),\Theta_f). \tag{2}$$

Here, c is the d-dimensional R-CNN descriptor [16] encoding the visual appearance of the foreground region in x. In the experiments, $d = 4096$. $\bar{x} = 1 - x$ is the complementary matrix of x, which specifies the background region. $f(\cdot)$ is the foreground discriminant function, which is a linear regressor parameterized by $\Theta_f = (w, b)$. The divergence term increases either: (1) the foreground region is more consistent with the foreground model; or (2) the background region is more different with the foreground model.

The prior term $P(x|I)$ evaluates the foreground segmentation x based on the low-level image features, namely, the color distribution, the boundary response and the visual saliency. It it defined as

$$P(x|I) = \|\boldsymbol{h}(x|I) - \boldsymbol{h}(\bar{x}|I)\|_1 + \lambda_1 B(x|I) + \lambda_2 S(x|I). \tag{3}$$

The first term measures the difference of the foreground region and the background region [17], where $\boldsymbol{h}(x|I)$ is the color histogram of mask x. The second term $B(x|I)$ measures the boundary strength of x using generalized boundary detection [18]. Finally, the third term measures the visual saliency of the foreground region using [19]. λ_1 and λ_2 are the weights, which can be obtained by grid search.

To compute the optimal labeling of the frame, we use parametric min-cut [20] to generate a pool of segmentation hypotheses $\mathcal{H}_t = \{x_i^t\}_{i=1}^{|\mathcal{H}_t|}$, where $|\mathcal{H}_t| = 50$ in the experiments. The edge weights between pixel nodes are calculated by [18] with both the color layer and the optical flow layer. Finally, the optimal segmentation of the frame is

$$x^\star = \underset{x\in\mathcal{H}}{\operatorname{argmax}}\, E(x|I,\Theta_f). \tag{4}$$

which can be computed efficiently, e.g., by greedy search.

3. MULTI-COMPONENT TRANSDUCTIVE LEARNING ON THE TEMPORAL TREES

To capture the foreground object which exhibits large variations in visual appearance in the video, a multi-component foreground model is proposed, which is parameterized by $\Theta_f = \{w_k, b_k\}_{k=1}^M$, where M is the number of components. Without loss of generality, we assume that M is a constant. Each component is parameterized by $w_k \in \mathbb{R}^d$ and $b_k \in \mathbb{R}$, which are the coefficients of the linear predictor. Thus, given the feature vector c of the foreground region, the discriminant function of the k-th component is

$$f(c, w_k, b_k) = \langle c, w_k \rangle + b_k, \quad k = 1, \cdots, M. \tag{5}$$

It measures the compatibility of the foreground region of the frame with the i-th component model.

To fit the multi-component foreground model Θ_f, a transductive learning algorithm is proposed, which "transfers" the information from the labeled frames to the unlabeled frames along a tree-structured graphical model, named *temporal tree*. Each node in the temporal tree is a frame, and each branch from the root to the leaf consists of visually similar and temporally coherent frames, which defines a component. The construction of the temporal tree is shown in Algorithm 1. The roots of the temporal tree are initialized with the labeled frames, and the unlabeled frames are added to the temporal tree recursively. The construction of the temporal tree is completed until all frames are added to the tree.

Here, we describe how the temporal tree grows. Let $\mathcal{L}$ be the set of the leaves of the current tree (starting from the labeled frames). The frames along the path from each leaf $V_k \in \mathcal{L}$ to the root is used to train a foreground component (w_k, b_k) by support vector regression. To be specific, a foreground regressor predicts the overlap of a foreground hypothesis with the ground truth foreground. Given the training set $\{c_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^m$, where m is the number of samples, c_i is the descriptor of the i-th hypothesis, and $y_i \in [0, 1]$ is the intersection-over-union ratio of the i-th hypothesis and the ground truth foreground. The optimal parameters of the component model can be computed by solving

$$\begin{aligned} &\min_{w_k, b_k} \frac{1}{2}\|w_k\|^2 + C\sum_{i=1}^m (\xi_i + \xi_i^*) \\ &\text{s.t.} \begin{cases} y_i - \langle w_k, c_i \rangle - b_k \le \epsilon + \xi_i, i = 1, \cdots, m \\ \langle w_k, c_i \rangle + b_k - y_i \le \epsilon + \xi_i^*, i = 1, \cdots, m \\ \xi_i \ge 0, i = 1, \cdots, m \\ \xi_i^* \ge 0, i = 1, \cdots, m \end{cases} \end{aligned} \tag{6}$$

where ϵ is the deviation and C is the trade-off parameter. The optimal w_k and b_k can be computed by solving the dual formulation of Eq. (6) via quadratic programming.

With this component model, the temporal tree grows by adding some unlabeled frames in the temporal neighborhood of the frame at V_k to the children of V_k. Let S be the degree of the temporal tree, *i.e.*, the maximum number of children that a node can have. The compatibility of a segmentation hypothesis with the k-th component model is measured by the discriminant value of the linear regressor of the component model as Eq. (5). Finally, the top S unlabeled frames in the temporal neighborhood of V_k with largest discriminant values are added to the temporal tree as the children of V_k. In this way, the temporal tree grows by iteratively adding the unlabeled frames to the leaves of the tree, until all frames are added. Overall, the temporal tree can be regarded as the extension to the bootstrapping transductive learning with multiple components.

Algorithm 1: Construction of the temporal tree

Input: Video frames: $\{I_k\}_{k=1}^N$;
Segment hypotheses: $\{x_k\}_{k=1}^N$
Output: Temporal tree
Initialize the temporal tree with the labeled frames;
while *not all frames are added the tree* **do**
  Collect the leaf set $\mathcal{L}$ of the current tree;
  for *each leaf* $V_k \in \mathcal{L}$ **do**
    Train an SVR from V_k to its root;
    Find the unlabeled temporal neighbors of V_k: $\mathcal{N}_k$;
    Compute the discriminant values of the frames in $\mathcal{N}_k$ by Eq. (5);
    Add S frames in $\mathcal{N}_k$ with the largest discriminant values to the children of V_k;
  end
end

4. EXPERIMENTS

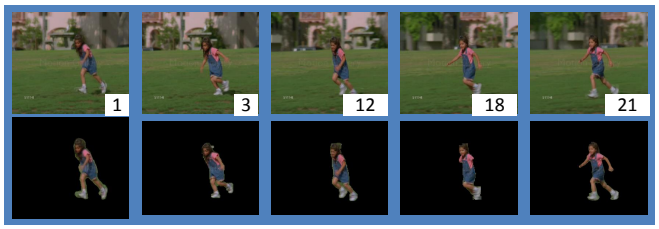
The experiments are conducted upon the SegTrack dataset [25] with six challenging video sequences, whose first frame is manually labeled. The proposed method is compared with four state-of-the-art video segmentation methods: (1) multi-class cosegmentation [21], (2) level-set based video segmentation [22], (3) graph based video segmentation [23], and (4) key-segments based video segmentation [24]. For [21], we change the class number from 2 to 9, and select the best segmentation result. The *mean pixel error* [24, 25] is used to evaluate the accuracy of video segmentation, which is defined as

$$e = \frac{1}{N}\sum_{t=1}^N XOR(x_t, x_t^{gt}), \tag{7}$$

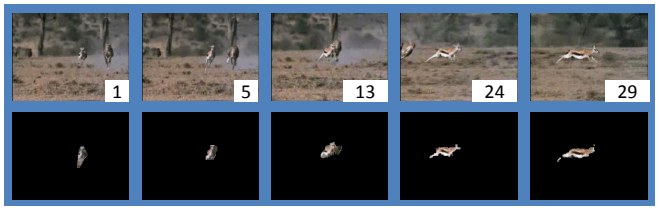
where x_t^{gt} is the ground truth foreground mask of the t-th frame. The quantitative result is shown in Table 1, and some segmentation examples are displayed in Fig. 2.

Table 1. Quantitative results and comparison on the SegTrack dataset

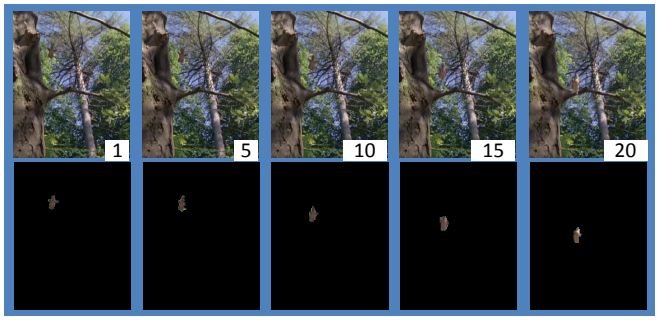
	birdfall	*cheetah*	*girl*	*monkeydog*	*parachute*	*penguin*
Proposed	**190**	**753**	1871	722	387	4841
multi-class cosegmentation [21]	988	3279	5321	1125	3245	8932
level-set based video segmentation [22]	454	1217	**1755**	683	502	6627
graph based video segmentation [23]	305	1219	5777	**493**	1202	**2116**
key-segments based video segmentation [24]	288	905	1785	493	**201**	136285



(a) Girl



(b) Cheetah



(c) Birdfall

Fig. 2. The segmentation results of the *Girl*, *Cheetah* and *birdfall* sequence in the SegTrack dataset.

The proposed method outperforms multi-class cosegmentation [21] on all the test sequences. Compared with level-set based video segmentation [22] and graph based video segmentation [23], the proposed method obtains higher accuracy in 4 out of 6 sequences, and the mean pixel error is reduced by 22% and 21% on average for all the sequences. Compared with key-segments based video segmentation [24], the proposed method obtains higher accuracy in 3 out of 6 sequences.

To further evaluate the performance of the proposed algorithm for objects with fast motion, the *Monkeydog* sequence is down-sampled by the ratio of 0.1 in the temporal domain to generate a new sequence, where the foreground object has extremely large motion. The segmentation result is displayed in Fig. 3. Clearly, the proposed method produces accurate segmentation of the foreground object with large motion. In comparison, optical flow based approach [23] gives different labels to the same object.

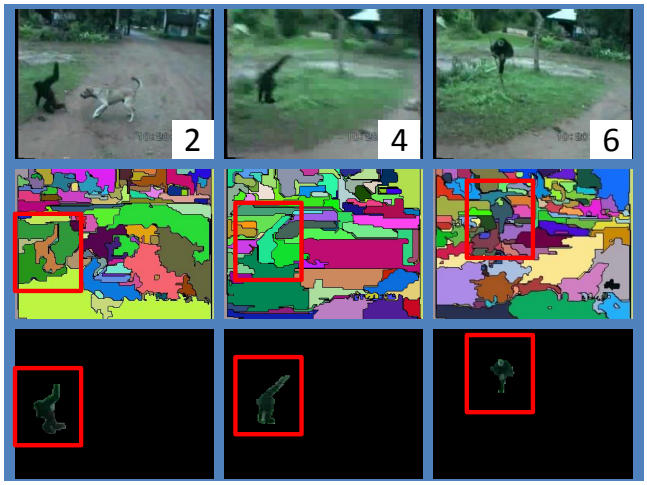


Fig. 3. Sample results from the temporally down-sampled *Monkeydog* sequence. The results of [23] are shown in the second row. The results of the proposed method are shown in the last row.

The experiments are conducted on a computer with Intel Core i7-3770 CPU and 16GB RAM, and the average run-time over the testing videos is about 4 minutes per frame.

5. CONCLUSION

In this paper, we proposed a multi-component transductive video cosegmentation approach. It involves a temporal coherent video co-segmentation framework, where the foreground object is modeled by multiple foreground components. To train the parameters in the model, a transductive support vector regression algorithm is performed in the temporal tree, which automatically collects visually and temporally consistent frames along the branches. Experimental results show that the proposed method outperforms many state-of-the-art approaches on public benchmark.

References

[1] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun, "Bottom-up segmentation for top-down detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 3294–3301.

[2] S. Duffner, S. Duffner, and C. Garcia, "Pixeltrack: A fast adaptive algorithm for tracking non-rigid objects," in *Proc. Int'l Conf. Computer Vision*, Sydney, NSW, Australia, Dec 2013, pp. 2480–2487.

[3] M. Hoai, Z. Lan, and F. Torre, "Joint segmentation and classification of human actions in video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2011, pp. 3265–3272.

[4] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, "Track to the future: Spatio-temporal video segmentation with long-range motion cues," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2011.

[5] Y. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proc. Int'l Conf. Computer Vision*, Barcelona, Spain, Nov 2011, pp. 1995–2002.

[6] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Tran. Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1187–1200, June 2014.

[7] M. Narayana, A. Hanson, and E. Learned-Miller, "Coherent motion segmentation in moving camera videos using optical flow orientations," in *Proc. Int'l Conf. Computer Vision*, Sydney, NSW, Australia, Dec 2013, pp. 1577–1584.

[8] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *Proc. Int'l Conf. Computer Vision*, Kyoto, Japan, Sept 2009, pp. 1219–1225.

[9] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar, "Weakly supervised learning of object segmentations from web-scale video," in *ECCV*, Florence, Italy, Oct 2012, pp. 198–208.

[10] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, "Discriminative segment annotation in weakly labeled video," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 2483–2490.

[11] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010, pp. 1943–1950.

[12] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010, pp. 3169–3176.

[13] W. Chiu and Mario Fritz, "Multi-class video co-segmentation with a generative multi-video model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 321–328.

[14] J. Rubio, J. Serrat, and A. Lopez, "Video cosegmentation," in *ACCV*, Daejeon, Korea, Nov 2012, pp. 13–24.

[15] D. Chen, H. Chen, and L. Chang, "Video object cosegmentation," in *Proc. ACM Multimedia*, New York, NY, USA, 2012, pp. 805–808.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 580–587.

[17] M. Tang, L. Gorelick, O. Veksler, and Y. Boykov, "Grabcut in one cut," in *Proc. Int'l Conf. Computer Vision*, Sydney, NSW, Australia, Dec 2013, pp. 1769–1776.

[18] M. Leordeanu, R. Sukthankar, and C. Sminchisescu, "Generalized boundaries from multiple image interpretations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1312–1324, July 2014.

[19] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vision*, vol. 13, no. 4, pp. 1–20, March 2013.

[20] J. Carreira and C. Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric mincuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1312–1328, July 2012.

[21] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012, pp. 542–549.

[22] P. Chockalingam, N. Pradeep, and S. Birchfield, "Adaptive fragments-based tracking of non-rigid objects using level sets," in *Proc. Int'l Conf. Computer Vision*, Kyoto, Japan, Sept 2009, pp. 1530–1537.

[23] M. Grundman, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010, pp. 2141–2148.

[24] Y. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proc. Int'l Conf. Computer Vision*, Barcelona, Spain, Nov 2011, pp. 1995–2002.

[25] David Tsai, Matthew Flagg, Atsushi Nakazawa, and JamesM. Rehg, "Motion coherent tracking using multi-label mrf optimization," *Int'l J. Computer Vision*, vol. 100, no. 2, pp. 190–202, 2012.