FIGURE/GROUND VIDEO SEGMENTATION USING GREEDY TRANSDUCTIVE COSEGMENTATION

Zhihui Fu, Hongkai Xiong

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

ABSTRACT

Cosegmentation has achieved great success in exploiting inter-image segmentation consistency to segment a group of images simultaneously. To enforces non-local temporal coherence across all the frames by high-order object-level appearance/semantic correspondence with a compensation to the short-time window motion coherence cue, this paper cosegments the video frames together with a novel inter-frame segmentation consistency term. A direct application of existing cosegmentation algorithms to video frames encounters the following challenges: the high correlation of adjacent frames which makes the segmentation ambiguous and a large number of video frames which makes the computation expensive. To tackle them, we formulate the cosegmentation in a transductive learning framework to iteratively learn the inter-frame consistency term from all the video frames. The proposed algorithm is evaluated on the standard SegTrack dataset and promising results are obtained.

Index Terms— video co-segmentation, transductive co-segmentation, greedy transductive inference, parametric min-cut, figure-ground video segmentation

1. INTRODUCTION

Figure-ground video segmentation is a classical and fundamental problem in computer vision. Many video segmentation approach base on the spatial-temporal graph-cuts [1]. The temporal coherence of the segmentation is enforced by linking temporal adjacent pixels by motion in a short-time window, which is generally good, but at cases when the motion field is noisy, the temporal coherence may not hold well. High-order Markov Random Field (MRF) is used to incorporate motion coherence [2] and shape [3]. In this paper, we would like to leverage object-level appearance/semantic consistency in all the video frames to compensate the motion cue for maintaining temporal consistency in the figure/ground video segmentation.

Cosegmentation has been actively studied in recent years. It aims to simultaneously segment a group of images, with the coherence of segmentations maintained across all the images in the group, where the coherence may refer to being the same object or being in the same (object) class. The typical cosegmentation approaches involve minimizing the MRF energy and the foreground histogram distances measured by L_1 norm [4], reward model [5] and the Boykov-Jolly model [6]. In 2012, an iterative scheme deals with multiple foregrounds of interest by alternating between a foreground modeling module and a region assignment module [7]. Lately, an energy-maximization cosegmentation approach [8] that can handle multiple classes through combining spectral- and discriminative-clustering terms and optimize the energy function using an efficient expectation-minimization (EM) algorithm. Illuminated by the insight, this paper focuses on figure-ground segmentation using temporal coherence and makes an inspiring solution to the challenges: the highly correlated background and the scalability to large number of video frames.

Initially, the video frames are cosegmented with an inter-frame coherence term that explicitly combines the high-order object-level appearance/semantic coherence, e.g. the bag of words object detector, and the temporal smoothness constraint based on motion. With the inter-frame coherence term, it can handle situations when the motion field is noisy caused by large motion and highly deformable shapes. To deal with the ambiguity of cosegmentation from the highly correlated backgrounds, the inter-frame coherence term is learned on all the video frames in a transductive framework with a greedy EM-like solution.

The rest of this paper is organized as follows. Section 2.1 defines the energy function, and Section 2.2 describes the transductive cosegmentation framework. The experimental results on SegTrack are evaluated in Section 3. Finally, we conclude this paper and discuss the future work in Section 4.

2. METHODOLOGY

Let k donate the video frame index and $i \in \{1, \ldots, n\}$ range over pixels in each frame.

- $x_k(i) \in \{0, 1\}$ donates if the pixel i in frame k belongs to the foreground. $\mathbf{x}_k$ is the labeling of frame k and $\mathbf{X}$ is the labeling of the whole video.

The work was supported in part by the NSFC, under grants U1201255, 61271218, and 61228101.

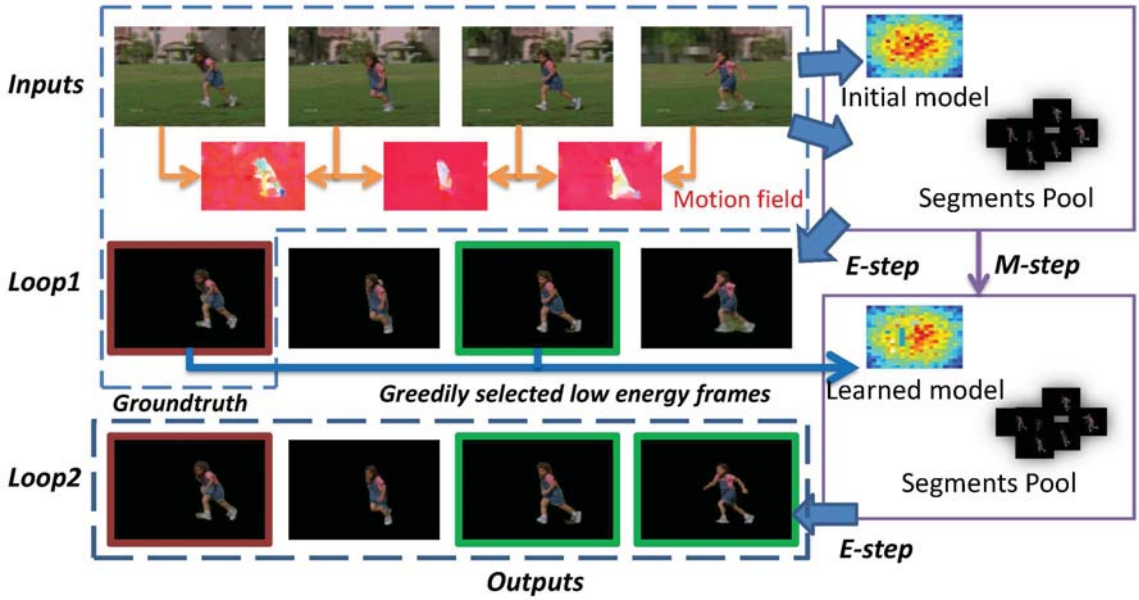

Fig. 1. The diagram of the transductive video cosegmentation framework, where the inputs are the video frames with one frame labeled (the frame with a brown border). The initial model is extracted from the labeled frame and has an impact on all the frames using parametric min-cuts. The frames with the lowest energies (in green borders) are used to train the model in an iterative EM-like solution.

- $z_k(i)$ is the measurement. $\mathbf{z}_k$ is shorthand for frame k and $\mathbf{Z}$ is the measurement of the whole video.

- $\theta_f(k)$ is the foreground model parameters for frame k. $\theta_b(k)$ is the background model parameters. We define $\theta_k = (\theta_f(k), \theta_b(k))$ and $\mathbf{\Theta}$ for the whole video.

Given the video frames I_i, the goal is to segment the frames simultaneously and to obtain the pixel labeling $x_i \in \{0, 1\}$ in each frame. In other words, it would enforce the nonlocal temporal consistency of the segmentations across all the video frames with high-order object-level appearance and semantic correspondence in addition to motion estimation. Specifically, the correspondence is measured by a learned color model and a bag of words detector. It is worth mentioning that any appearance feature or the object detector can be used, e.g. the deformable part models [9]. To deal with the problem from the high correlation among the frame backgrounds, a greedy transductive cosegmentation is proposed, which is depicted in Fig. 1.

2.1. Energy function

Inspired by cosegmentation, a general model $\mathbf{\Theta}$ from all the video frames is constructed to constraint the segmentation consistency that captures the video foreground and background structures. Without loss of generality, the general model is represented in a nonparametric form: a normalized rgb color histogram model $\mathbf{h}_l^c, l = \{f, b\}$ and a bag of words detector with histograms $\mathbf{h}_l^w, l = \{f, b\}$ to model the foreground object as well as the video background, i.e. $\mathbf{\Theta} = (\mathbf{h}_l^c, \mathbf{h}_l^w)$. To generate the bag of words model $\mathbf{h}_l^w$, dense color SIFT descriptors are extracted at a grid on the video frames. All the extracted descriptors are clustered using the K-means algorithm to construct a codebook of visual vocabularies. In turn, the descriptors in each frame are quantified according to the codebook and the words located in the object foreground and background are, respectively, replenished to update the foreground bag of words histogram and the background bag of words histogram. The models are expressed as $\mathbf{h}_l^u$ with $l = \{f, b\}$ and the histograms extracted in each frame $\mathbf{z}_k$ with a labeling $\hat{\mathbf{x}}_k$ are donated as $\mathbf{h}_{lk}^u(\hat{\mathbf{x}}_k)$. Hence, the object coherence is defined on the distance metric between the model representations. To be concrete, we define the metric as a linear combination of the χ^2 kernels.

The energy function is shown in Eq. 1:

$$E(\mathbf{X}, \mathbf{\Theta}) = \sum_k S(\mathbf{x}_k, \mathbf{\Theta}) + C(\mathbf{x}_k, \mathbf{\Theta}) \tag{1}$$

where the first term $S(\mathbf{x}_k, \mathbf{\Theta}) = -\|\mathbf{h}_{fk}^c(\hat{\mathbf{x}}_k) - \mathbf{h}_{bk}^c(\hat{\mathbf{x}}_k)\|_{L_1}$ is the intra frame segmentation term as [10]. The second term is the inter-frame coherence term that encourages the temporal consistency of the segmentations across all the frames.

$$C(\mathbf{x}_k, \mathbf{\Theta}) = \sum_{u=w,c} \sum_{l=f,b} \alpha_{ul} \kappa(\mathbf{h}_l^u, \mathbf{h}_{lk}^u(\hat{\mathbf{x}}_k)) - \lambda m(\mathbf{x}_k, \mathbf{s}_k) \tag{2}$$

where $\kappa(\mathbf{x}, \mathbf{y}) = \sum_i \frac{2x_i y_i}{x_i + y_i}$ is the χ^2 kernel, $\mathbf{s}_k$ encodes the

motion information by propagating the segmentation mask from the previous frame to the current frame. Function $m(\cdot)$ measures the affinity of two masks.

The first term in Equation 2 exploits the inter-frame segmentation consistency, which is defined by high-order object-level features as in [7] and enforced on all the video frames. When $m(\cdot)$ is adopted in an overlap manner, the second term in Eq. 2 is equivalent to a conventional graph-cuts temporal smoothness term. Overall, the proposed temporal coherence term is an extension of the motion-based temporal smoothness as a combination of the object-level appearance/semantic overlap and the motion-based temporal smoothness. Since $\mathbf{\Theta}$ is not known in advance, $C(\mathbf{X}, \mathbf{\Theta})$ should be learned from all the video frames in a transductive learning framework in Section 2.2.

2.2. Transductive learning with expectation maximization

Recognizing that the video frames usually shares similar backgrounds, it is reasonable to exploit cosegmentation by learning the inter-frame consistency term on the high-correlated video frames. To deal with the ambiguity, we leverage the segmentation $\mathbf{x}_e$ of one frame $\mathbf{z}_e$ to give an initial hint on the appearance of the foreground object.

The learner in the transductive learning setting [11], in addition to labeled training samples, has access to all of the unlabeled test samples. The objective is to determine the labels of the test samples by learning on both the training samples and the test samples. The transductive inference can use the structure and distributions of the test samples, and have achieved success when there is a small number of training instances and a large number of test instances.

Given all the video data $\mathbf{Z}$ along with one labeled frame $\mathbf{z}_e$, $\mathbf{\Theta}$ would be learned by the transductive learning to exploit the information hidden in the unlabeled frames. Essentially, learning is guided by the object function

$$F(\mathbf{\Theta}, \mathbf{X}) = \lambda_s \ell(\mathbf{\Theta}, \mathbf{x}_e) + \lambda_t \sum_{k \neq e} \ell(\mathbf{\Theta}, \mathbf{x}_k) \qquad (3)$$

where $\ell(x, y)$ is the loss function defined by Eq. 1

$$\ell(\mathbf{\Theta}, \mathbf{x}_k) = S(\mathbf{x}_k, \mathbf{\Theta}) + C(\mathbf{x}_k, \mathbf{\Theta}) \qquad (4)$$

Eq. 3 minimizes the model losses on the training sample as well as on the test video frames.

For a small number of test instances, we can attempt all the possible configurations. However, the data volume of a typical video is large. In practice, we approximate the transductive cosegmentation problem in a greedy EM-like optimization. With fixed iteration times, the computation complexity grows linearly with the number of video frames.

2.2.1. E-step

Given a fixed $\mathbf{\Theta}$, the energy function in Eq. 1 is minimized is minimized within a pool of segment proposals p_k generated for each frame $\mathbf{z}_k$. A parametric min-cut algorithm similar to [12] is utilized to construct the proposals. The parametric min-cut can be casted as a minimization problem:

$$\min_{x_i=\{0,1\}} \sum_{x_i} (a_i + \lambda_i b_i) x_i + \sum_{x_i \ x_j} \delta(x_i, x_j) g(x_i, x_j) \qquad (5)$$

where foreground pixels are labeled with 1, and if x_i and x_j are adjacent, $x_i \ x_j$. $\delta(x_i, x_j)$ takes 1 if x_i and x_j are not equal. $g(x_i, x_j)$ measures similarity between pixels and is calculated via gPb [13]. The parametric min-cut differs to the conventional min-cut algorithms in that the output segment scale would be adjusted by the parameters λ_i. With a large λ_i,, the output segment scale is small, and vice versa. In fact, regions with motion different from its surroundings are likely to be belong to the foreground object. Hence, we apply parametric min-cut on the optical flow field to generate the proposals with weak color edges but a strong motion boundary. The generated segment pool consists hundreds of proposals and covers the foreground object and its parts, which could provide a much smaller search space than the original pixel-wise space. The proposal in p_k with the lowest energy will be chosen as the segmentation for frame $\mathbf{z}_k$.

2.2.2. M-step

Given a segmentation $\hat{\mathbf{X}}$ of all the video frames, $\mathbf{\Theta}$ is updated by Eq. 3. Specifically, the N lowest energy segmentations $\hat{\mathbf{x}}_s$ with $s \in \{1, \ldots, N\}$ and weighted average $\mathbf{h}^u_{ls}(\hat{\mathbf{x}}_k)$ with $\mathbf{h}^u_l$ would be selected.

3. EXPERIMENTAL RESULTS

We adopt the SegTrack dataset [2] to evaluate the performance of the proposed transductive video cosegmentation approach. The SegTrack dataset composes six sequences that feature video segmentation challenges: color overlap, inter-frame motion and shape deformation. The optical flow is obtained by the Classic+NL algorithm [14]. Compared to the up-to-date level-set approach in [15], we evaluated the

Table 1. Quantitative results and comparison on the SegTrack dataset

	Proposed	only motion	[15]	CPMC
birdfall	911	1311	454	869
cheetah	1321	1335	1217	657
girl	1850	1957	1755	1820
monkeydog	571	652	683	489
parachute	490	477	502	317
penguin	5372	6713	6627	4085

Fig. 2. The sampled results from the *Girl* sequence, the *Monkeydog* sequence and the *Cheetah* sequence of the SegTrack dataset. The original frames are shown in row 1, 3, 5, and the segmentation results are shown in row 2, 4, 6.

segmentation quality in Table 1 using per-frame pixel error mean $e = \frac{XOR(x,gt)}{n}$, where n is the number of frames. The results are listed in Table 1. Since the proposed segmentation is operated based on the segments provided by CPMC scheme [12], its performance is up-bounded by the best CPMC results which are derived from the groundtruth masks to select the best proposal from the segments pool. Obviously, the proposed effect is close to the upper bound. The "only motion" column is obtained from set α_{ul} to zero in Eq. 2, and generally behaves worse than the proposed scheme, which illustrates the effect of the high-order coherence term. In particular, the algorithm performs well on the *Monkeydog* sequence where the foreground involves intensive movements. More results are illustrated in Fig. 2. To further depict the proposed algorithm's capability in segmenting foreground objects with severe motion, we sampled three frames from the *Monkeydog* sequence where the foreground object has moved a large distance. Acting the proposed scheme on the three frames, Fig. 3 shows that the foreground object is correctly cut out.

4. CONCLUSION

In this paper, we proposed a figure/ground video segmentation algorithm using a greedy transductive cosegmentation on all the video frames. A novel inter-frame segmentation consistency term combining the high-order appearance/semantic models with the motion, is proposed to handle scenes with noisy motion field. To tackle the ambiguity from the highly correlated frame appearance, the inter-frame term is learned with the help of a labeled frame in a transductive inference framework with a greedy EM-like solution. The future work would involve an automatic tuning of parameters so that it can be more flexible to fit the properties of various video segmentation scenes.

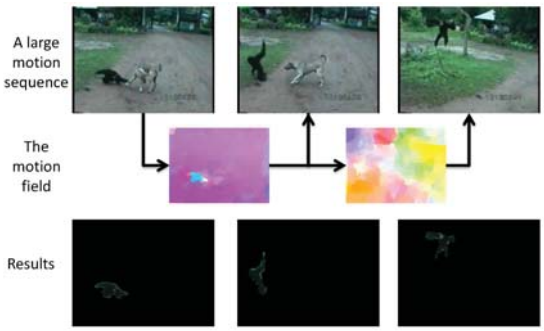

Fig. 3. Taking three sampled frames from the *Monkeydog* sequence, where the foreground object "Monkey" involves large motion and the motion field is distorted for frame 2 to frame 3. With the proposed inter-frame segmentation consistency term, the foreground object is correctly segmented.

5. REFERENCES

[1] J. Malcolm, Y. Rathi, and A. Tannenbaum, "Multi-object tracking through clutter using graph cuts," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–5.

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

[3] Chun hao Wang and Ling Guan, "Graph cut video object segmentation using histogram of oriented gradients," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, 2008, pp. 2590–2593.

[4] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, vol. 1, pp. 993–1000.

[5] D.S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 269–276.

[6] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother, "Cosegmentation revisited: Models and optimization," in *Computer Vision - ECCV 2010*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios, Eds., vol. 6312 of *Lecture Notes in Computer Science*, pp. 465–479. Springer Berlin Heidelberg, 2010.

[7] Gunhee Kim and Eric P. Xing, "On Multiple Foreground Cosegmentation," in *25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, 2012.

[8] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 542–549.

[9] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[10] Meng Tang, Lena Gorelick, Olga Veksler, and Yuri Boykov, "Grabcut in one cut," in *International Conference on Computer Vision (ICCV)*, Sydney, Australia, December 2013.

[11] Thorsten Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the Sixteenth International Conference on Machine Learning*, San Francisco, CA, USA, 1999, ICML '99, pp. 200–209, Morgan Kaufmann Publishers Inc.

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

[13] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[14] Deqing Sun, S. Roth, and M.J. Black, "Secrets of optical flow estimation and their principles," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2432–2439.

[15] P. Chockalingam, N. Pradeep, and S. Birchfield, "Adaptive fragments-based tracking of non-rigid objects using level sets," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 1530–1537.