Graph Transduction Learning of Object Proposals for Video Object Segmentation

Tinghuai Wang[⋆] and Huiling Wang[⋆]

Nokia Technologies, Tampere, Finland

Abstract. We propose an unsupervised video object segmentation algorithm that detects recurring objects and learns cohort object proposals over space-time. Our core contribution is a graph transduction process that learns object proposals densely over space-time, exploiting both appearance models learned from rudimentary detections of sparse object-like regions, and their intrinsic structures. Our approach exploits the fact that rudimentary detections of recurring objects in video, despite appearance variation and sporadity of detection, collectively describe the primary object. By learning a holistic model given a small set of object-like regions, we propagate this prior knowledge of the recurring primary object to the rest of the video to generate a diverse set of object proposals in all frames, incorporating both spatial and temporal cues. This set of rich descriptions underpins a robust object segmentation method against the changes in appearance, shape and occlusion in natural videos.

1 Introduction

Video segmentation remains an open challenge for Computer Vision, with recent advances relying upon prior knowledge supplied via interactive initialization or correction [1–6]. Yet fully unsupervised video segmentation [7–11] remains useful in Big Data scenarios such as video summarization or ingest pre-processing for video indexing or recognition, where the human in the loop is impractical. This is a very challenging task due to the lack of prior knowledge about object appearance, shape or position. Furthermore, variance in illumination and occlusion relationships introduce ambiguities that in turn induce instability in boundaries and the potential for localized under- or over-segmentation.

This paper proposes a novel automatic video object segmentation algorithm in which the segmentation of each frame is driven by set of rich object models learned from *spatio-temporally dense and coherent object proposals*. The core novel contribution is our *graph transduction* approach to the efficient learning of the dense video object proposals which enables the detection and segmentation of objects in complex dynamic scenes without suffering from appearance variation or object occlusion over time. In contrast to previous techniques, our algorithm learns and extracts object proposals from scratch to account for the evolution of object's appearance, shape and location with time, as opposed to selecting from existing per-frame detections of object-like regions [9–12].

[⋆] indicates equal contribution

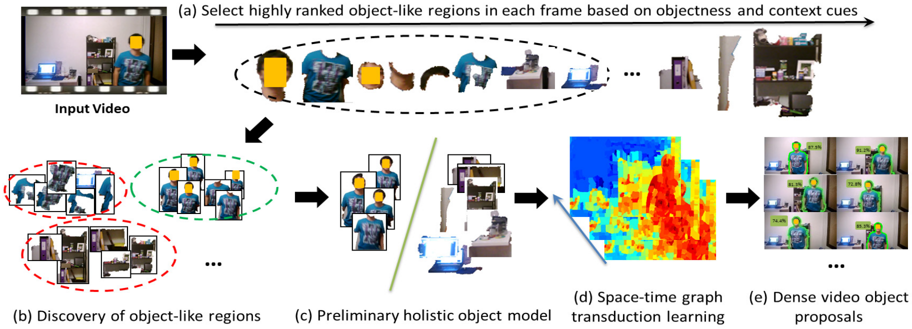
(a) Select highly ranked object-like regions in each frame based on objectness and context cues
Input Video
...
...
(b) Discovery of object-like regions
(c) Preliminary holistic object model
(d) Space-time graph transduction learning
...
(e) Dense video object proposals

Fig. 1. Generation of dense video object proposals.

The key idea is to create feature-based rudimentary detections of regions for the primary object by discriminative learning from labelled examples of sparse object-like regions. These detections serve as informative indicators of the appearance and location of the object. We propagate this labeled data on an undirected space-time graph consisting of regions, solving the graph transduction learning efficiently with a fast convergence technique [13]. Inference at the region level further makes our dense video object proposal extraction approach a practical solution for unsupervised object segmentation on natural video sequences.

2 Related Work

Video object segmentation methods requiring user to provide an initial annotation of the first frame have been proposed, which either propagate the annotation to drive the segmentation in successive frames [1–6] or perform spatio-temporal grouping [14, 15]. The former group of methods heavily rely on motion estimates and may fail in segmenting videos with complex motions or varying object appearance. Although stability is achieved in the latter methods, they usually become computationally infeasible for pixel counts in even moderate size videos, and often fail in dealing with fast moving objects.

Automatic or unsupervised methods have also been proposed as a consequence of the prohibitive cost of user intervention in processing large amounts of video data in most computer vision applications. Methods like [16–21] achieve segmentation in a bottom-up approach based on spatio-temporal appearance and motion constraints. Motion segmentation methods cluster pixels or superpixels in video employing long-term motion trajectories analysis, which require the motion of the primary object to be neither too similar with the background nor too fast. Methods which generate over-segmentations for later processing analog to still-image superpixels [22] have also been proposed [23, 24], by applying spatio-temporal clustering based on low level features. However, without any top-down explicit notion of object, all of these automatic methods produce segmentations without corresponding to any particular object with semantic meaning.

Several recent methods [9–12] are proposed based on exploring recurring object-like regions from still images by measuring generic object appearance [25]. Lee *et al.* [9] proposed to extract 'key-segments' of the primary object by performing clustering in a pool of object proposals from each frame of the video. The weakness of this approach is that the object proposal pool combines regions across all frames and discards the spatial and temporal information of each region. Ma and Latecki [10] proposed to leverage the temporal information by utilizing binary appearance relation between regions in different frames and model the object region selection as a constrained Maximum Weight Cliques problem. Zhang *et al.* [11] improved this approach by introducing optical flow to track the evolution of object shape and appearance and solving the primary object proposal selection problem as the longest path problem for Directed Acyclic Graph (DAG). There are mainly two limitations with these later two approaches [10, 11]. First, both approaches propose to select or merge per-frame extracted object-like regions based on the objectness score which is computed locally in each frame, regardless of the prior knowledge of the corresponding object learned from other frames; their performance heavily relies on the quality of the initial rudimentary detection of object-like regions which is highly unreliable in practice. The initial object proposals generated using [25] normally contain a large amount of erroneous regions. Second, both approaches assume all object-like regions within each frame are independent and do not explicitly consider spatial affinity. This substantially limits the size of the object proposal especially when the primary object is comprised of multiple regions with distinct appearances. An additional limitation of [11] is that it employs optical flow warped region overlap to merge object-like regions into a new region which may introduce further spurious proposals due to inherent motion estimate error. Li *et al.* [12] proposed to track a pool of figure-ground segments in each frame and incrementally to learn a long-term object appearance model. However the incrementally built appearance model heavily relies on greedy matching and also suffers from the cumulative motion estimation error. All the above methods do not build an explicit holistic appearance model but relies on local heuristics and motion for selecting the object proposals.

To address the limitations of the above approaches [9–12], we propose to learn a holistic appearance model from the rudimentary detection of object-like regions across the whole video to drive the generation of dense object proposals. We propagate the prior knowledge from rudimentary detections on an undirected space-time graph consisting of regions by performing transduction learning, with respect to both low level cues collectively revealed by the appearance model and the intrinsic structure within video data. The transduction learning is guided by the initially detected evidence by collectively learning the initial sparse object-like regions, rather than directly using the local static 'objectness' score. Spatio-temporally coherent and dense object proposals are generated to facilitate robust object segmentation in challenging natural videos.

Our approach advances the state-of-the-art mainly in three aspects: (1) it explores the holistic patterns of primary object which are collectively revealed by a small set of object-like regions, and thus it prunes the spurious regions due to the independent rudimentary detection from a particular frame without considering the object-like regions generated in adjacent frames (2) it employs an efficient graph transduction learning approach to generating object proposals evenly and consistently distributed spanning the whole video, by exploiting both the local evidence and the intrinsic structure within video data (3) this set of object proposals provides sufficient and diverse appearance, shape, and location prior information to drive object segmentation while preserving spatio-temporal coherence.

3 Video Object Proposals

Our approach to generating video object proposals is comprised of three main steps: (1) object-like regions are extracted from each frame and a small set of the most likely object regions associated with the primary object in the video are identified (2) a holistic appearance model is learned from the object-like regions to describe the primary object spanning the whole video (3) in a top-down approach, transduction learning is performed on a space-time graph of regions to efficiently *generate* object proposals in each frame integrating the shared object models, temporal correlation and intrinsic structure within video data.

3.1 Initial Detection of Object-Like Regions

Since we assume no prior knowledge on the size, shape, appearance or location of the primary object, our algorithm operates by producing a diverse set of object proposals in each frames using [25] which is a category independent method to identify object-like regions in still image. To find the object-like regions among the proposals, we compute the 'objectness' of each region r as

$$S(r) = A(r) + M(r)$$

where $A(r)$ is the appearance score and $M(r)$ is the motion score. The static intra-frame appearance score $A(r)$ is computed using [25]. Motion score $M(r)$ reflects the disparity of motions between primary object and background. We compute optical flow [26] histograms for region r and $\bar{r}$ which is formed by merging all the closest surrounding regions of r. Using surrounding regions is more informative than using pixels in a loosely fit bounding box around r in [9]. We compute $M(r)$ as $M(r) = 1 - \exp(-\chi^2_{flow}(r, \bar{r}))$, where $\chi^2_{flow}(r, \bar{r})$ is the χ^2 distance between L_1-normalized optical flow histograms for regions r and $\bar{r}$.

Following [9], we firstly form a candidate pool $\mathcal{C}$ by taking the top N ($N = 10$) highest-scoring regions from each frame, and then identify groups of object-like regions that may represent a foreground object by performing spectral clustering in $\mathcal{C}$. All clusters are ranked based on the average score $S(r)$ of its comprising regions. The clusters among the highest ranks correspond to the most object-like regions but there may also be noisy regions, which is denoted as $\mathcal{H}$.

3.2 Holistic Appearance Model

Each object-like region from the rudimentary detection may correspond to different part of the primary object from particular frames, whereas they collectively describe the primary object. We could devise a discriminative model to learn the appearance of those most likely object regions. The initial set of object-like regions $\mathcal{H}$ form the set of all instances with a positive label (denoted as $\mathcal{P}$), while negative regions ($\mathcal{N}$) are randomly sampled outside the bounding box of the positive example. We use this labeled training set to learn linear SVM classifier for two categories. The classifier provides a confidence of class membership taking the features of a region which combines texture and color features, as input. This classifier is then applied to all the unlabeled regions across the whole video. After this classification process, each unlabelled region i is assigned with a weight Y_i from SVM, i.e. the signed distance to the decision boundary. All weights are normalized between -1 and 1, by the sum of unsigned distances to the decision boundary.

3.3 Graph Transduction Learning of Object Proposals

The holistic appearance model provides an informative yet independent and incoherent prediction on each of the unlabelled regions regardless the inherent structure revealed by both labeled and unlabeled regions. To generate robust dense video object proposals, we adopt a graph transduction learning approach, exploiting both the *intrinsic structure* within data and the *initial local evidence* from the holistic appearance model.

Space-Time Graph of Regions To perform transduction learning, we define a weighted space-time graph $\mathcal{G}_s = (\mathcal{V}, \mathcal{E})$ spanning the whole video with each node corresponding to a region, and each edge connecting two regions based on spatial and temporal adjacencies. Temporal adjacency is coarsely determined based on motion estimates. Each region r_i^k in frame i is warped by the forward optical flow to frame $i+1$ and the overlap ratio between the warped region r_i^k and the overlapped regions r_{i+1}^j in frame $i+1$ are computed as $S_{\text{overlap}}(k,j) = \frac{|\tilde{r}_i^k \cap r_{i+1}^j|}{|\tilde{r}_i^k|}$, where $\tilde{r}_i^k$ is the warped region of r_i^k by optical flow to frame $i+1$, and $|r|$ is the cardinality of region r. If $S_{\text{overlap}}(k,j)$ is greater than 0.5 for a pair of regions, i.e. r_i^k and r_{i+1}^j, in two successive frames, they are deemed temporally adjacent. Note that accurate motion estimation is neither assumed nor required to construct this graph.

We compute the affinity matrix W of the graph using the feature histogram representation h_{r_i} of each region r_i as $W_{ij} = \exp(-\frac{\chi^2(h_{r_i}, h_{r_j})}{2\beta})$, where β is the average χ^2 distance between all adjacent regions. Since sparsity is important to remove label noise and semi-supervised learning algorithms are more robust on sparse graphs [27], we set all W_{ij} are set to zero if r_i and r_j are not adjacent.

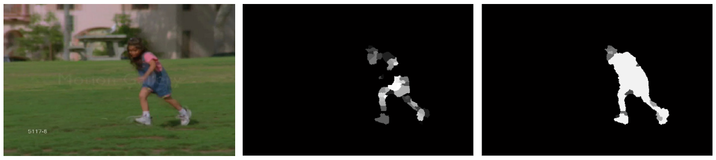

Fig. 2. Positive predictions of each region and the brightness indicates probability of being an object: (a) source image (b) independent SVM predictions (c) predictions from graph transduction capturing the coherent intrinsic structure within visual data, using SVM predictions as input.

Graph Transduction Learning Graph transduction learning propagates label information from labeled nodes to unlabeled nodes. Let the node degree matrix $D = \mathrm{diag}([d_1, \ldots, d_N])$ be defined as $D_i = \sum_{j=1}^{N} W_{ij}$, where $N = |\mathcal{V}|$. We follow a similar formulation with [13] to minimize an energy function $E(F)$ with respect to all region labels F ($F \in [-1, 1]$):

$$E(F) = \sum_{i,j=1}^{N} W_{ij} | \frac{F_i}{\sqrt{D_i}} - \frac{F_j}{\sqrt{D_j}} |^2 + \mu \sum_{i=1}^{N} |F_i - Y_i|^2, \tag{1}$$

where $\mu > 0$ is the regularization parameter, and Y are the desirable labels of nodes which are normally imposed by prior knowledge. The first term in (1) is the *smoothness constraint*, which encourages the coherence of labelling among adjacent nodes, whilst the second term is the *fitting constraint* which enforces the labelling to be similar with the initial label assignment.

The optimization problem in (1) is solved by an iteration algorithm in [13]. Alternatively we solve it as a linear system of equations. Differentiating $E(F)$ with respect to F we have

$$\nabla E(F)|_{F=F^*} = F^* - SF^* + \mu(F^* - Y) = 0 \tag{2}$$

where $S = D^{-1/2} W D^{-1/2}$. It can be transformed as

$$F^* - \frac{1}{1+\mu} SF^* - \frac{\mu}{1+\mu} Y = 0 \tag{3}$$

Denoting $\gamma = \frac{\mu}{1+\mu}$, we have $(I - (1-\gamma)S)F^* = \gamma Y$. An optimal solution for F can be solved using the Conjugate Gradient method with very fast convergence.

We use the predictions from SVM classifier to assign the values of Y. The diffusion process can be performed for positive and negative labels separately, with initial labels Y in (1) substituted as Y_+ and Y_- respectively:

$$Y_+ = \begin{cases} Y & \text{if } Y > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and

$$Y_- = \begin{cases} -Y & \text{if } Y < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Combining the diffusion processes of both the object-like regions and background can produce more efficient and coherent labelling, taking advantage of

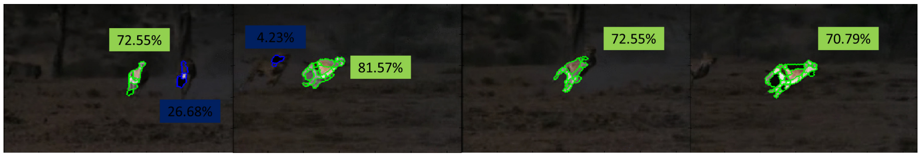

Fig. 3. Exemplar video object proposals from CHEETAH sequence. Colors of contour indicate different proposals. The transparency of each region indicates the objectness (F) from graph transduction learning. The objectness of each final object proposal is computed by averaging the constituent region-wise objectness F weighted by area.

their complementary properties. We perform the optimization for two diffusion processes simultaneously as follows:

$$F^* = \gamma(I - (1-\gamma)S)^{-1}(Y_+ - Y_-). \tag{6}$$

This enables a faster and stable optimization avoiding separate optimizations while giving equivalent results to the individual positive and negative label diffusion. Fig. 2 shows the positive predictions of each region, from SVM predictions and graph transduction learning respectively. The prediction from SVM exhibits unappealing incoherence, nonetheless, using it as initial input, graph transduction gives smooth predictions exploiting the inherent structure of data.

Finally, the regions which are assigned with label $F > 0$ from each frame are grouped. Specifically, we use the final label F to indicate the level of objectness of each region. The final proposals are generated by grouping the spatially adjacent regions ($F > 0$), and assigned by an objectness value by averaging the constituent region-wise objectness F weighted by area. The grouped regions with the highest objectness per frame are added to the set of object proposals $\mathcal{P}$. Exemplar video object proposals are shown in Fig. 3.

4 Video Object Segmentation

We formulate video object segmentation as a pixel-labelling problem of assigning each pixel with a binary value which represents background or foreground (object) respectively. We define a space-time graph by connecting frames temporally with optical flow displacement. In contrast to the previous space-time graph during transduction learning, each of the nodes in this graph is a pixel as opposed to a region, and edges are set to be the 4 spatial neighbors within the same frame and the 2 temporal neighbors in adjacent frames. We define the energy function that minimizes to achieve the optimal labeling:

$$E(x) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \lambda \sum_{i \in \mathcal{V}, j \in N_i} \psi_{i,j}(x_i, x_j)$$

where N_i is the set of pixels adjacent to pixel i in the graph and λ is a parameter.

The pairwise term $\psi_{i,j}(x_i, x_j)$ penalizes different labels assigned to adjacent pixels:

$$\psi_{i,j}(x_i, x_j) = [x_i \neq x_j]\mathrm{exp}(-d(x_i, x_j))$$

where $[\cdot]$ denotes the indicator function. The function $d(x_i, x_j)$ computes the color and edge distance between neighboring pixels:

$$d(x_i, x_j) = \beta(1 + |SE(x_i) - SE(x_j)|) \cdot ||c_i - c_j||^2$$

where $SE(x_i)$ ($SE(x_i) \in [0, 1]$) returns the edge probability provided by the Structured Edge (SE) detector [28], $||c_i - c_j||^2$ is the squared Euclidean distance between two adjacent pixels in CIE Lab colorspace, and $\beta = (2 < ||c_i - c_j||^2 >)^{-1}$ with $< \cdot >$ denoting the expectation.

The unary term $\psi_i(x_i)$ defines the cost of assigning label $x_i \in \{0, 1\}$ to pixel i, which is defined based on the per-pixel probability map by combining color distribution and region objectness:

$$\psi_i(x_i) = \begin{cases} -\log(w \cdot U_i^c(x_i) + (1 - w) \cdot U_i^o(x_i)) & \text{if } x_i \in \mathcal{P} \\ -\log U_i^c(x_i) & \text{otherwise} \end{cases} \tag{7}$$

where $U_i^c(\cdot)$ is the color likelihood and $U_i^o(\cdot)$ is the objectness cue. The definitions of these two terms are explained in detail next.

Color Likelihood

To model the appearance of the object and background, we estimate two Gaussian Mixture Models (GMM) in CIE Lab colorspace. Pixels belonging to the set of object proposals are used to train the GMM representing the primary object, whilst randomly sampled pixels in the complement of object proposals are adopted to train the GMM for the background. Given these GMM color models, per-pixel probability $U_i^c(\cdot)$ is defined as the likelihood observing each pixel as object or background respectively can be computed.

Objectness Cue

Extracted object proposals provide explicit information of how likely a region belongs to the primary object (objectness) which can be directly used to drive the final segmentation. Per-pixel likelihood $U_i^o(\cdot)$ is set to be the objectness value (F in (6)) of the region it belongs to.

Optimization

We adopt the binary graph cut [29] to minimize (7) and the resulting label assignment gives the foreground object segmentation of the video.

5 Implementation Details

We start by computing feature descriptors for all the regions in video. Two types of bag-of-features histograms are used: Texton Histograms (TH) and Color Histograms (CH). For TH, a filter bank with 18 bar and edge filters (6 orientations and 3 scales for each), 1 Gaussian and 1 Laplacian-of-Gaussian filters, is used.

400 textons are quantized via k-means. For CH, we use CIE Lab color space with 20 bins per channel (60 bins in total). All histograms are concatenated to form a single feature vector for each region. We learn 5 components per GMM to model the color distribution.

We empirically set $\mu = 3.0$ to balance the impact of the prior labelling and the local labelling smoothness. For graph cut optimization, we set $\lambda = 5$ and $w = 0.35$ by optimizing segmentation against ground truth over a training set of 5 videos which proved to be a versatile setting for a wide variety of videos. These parameters are fixed for the evaluation.

For efficiency and scalability, our region graph transduction learning is sequentially performed on clips of 20 frames by dividing the source video. The efficient transduction learning normally takes ~ 18 seconds on a clip of 20 frames with an unoptimized MATLAB implementation. The final graph cut based pixel labelling is sequentially performed in each frame in turn, using a space-time graph of three consecutive frames.

6 Experimental Results

We evaluate our method on two datasets: SegTrack [4] and a new dataset consisting of five videos. Two videos (*waterski, yunakim*) of this new dataset are from GaTech video segmentation dataset [19], two (*jump, gymnastic*) from the challenging VOT2013 [30] dataset, and one (*monkeybar*) from video tooning [14]. The SegTrack dataset comes with pixel-level ground truth for the task of video object segmentation. We manually labelled the ground-truth segmentation of all the frames in the new dataset for evaluation. We measure the segmentation performance as the average number of per-frame pixel error compared to the ground-truth, which is defined as [4] error $= \frac{\mathrm{XOR(S,GT)}}{\mathrm{NF}}$, where S denotes the label for every pixel in the video, GT is the ground-truth, and NF is the total number of frames in the video.

6.1 SegTrack Dataset

There are totally six videos (*birdfall, cheetah, girl, monekeydog, parachute, penguin*) in SegTrack dataset. We follow the setup in previous works [9–11, 21, 12] and discard the *penguin* video, since only a single penguin is labelled in the ground-truth amidst a group of penguins. Those videos exhibit a variety of challenges, including objects of similar color to the background, fast motion, non-rigid deformations, and fast camera motion.

Evaluation of Video Object Proposals To evaluate our method's capability to detect and generate spatio-temporal coherent and dense video object proposals, we firstly compare with [25], one of the state-of-the-art segment based object proposal methods on still images, as the baseline. Table 1 compares the per-pixel error rate of our object proposals, per-frame best scoring object proposal generated from [25], and also the lowest/highest error rates of all existing methods on SegTrack dataset. We observe that [25] returns inconsistent and sporadic object

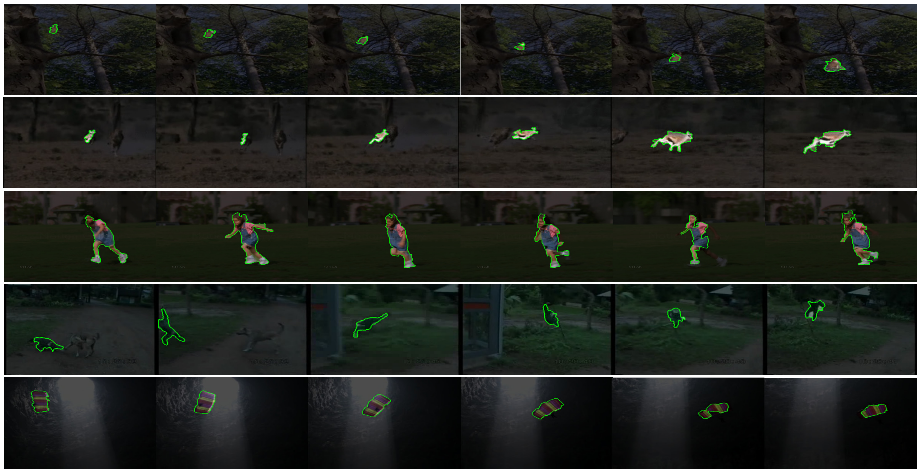

Fig. 4. Primary object proposals generated by the proposed graph transduction learning method.

Table 1. Quantitative results on SegTrack. The proposed video object proposals are compared with the per-frame top-scoring object proposal from [25], and also the lowest/highest error rates of all existing video object segmentation methods.

Video (No. frames)	Our Proposal	[25]	Lowest Error	Highest Error
birdfall (30)	264	22167	151	454
cheetah (29)	869	20649	633	1217
girl (21)	1683	8176	1121	1785
monkeydog (71)	839	29058	284	3859
parachute (51)	450	82934	201	855

proposals independently in each frame, whilst our object proposal captures the coherent essence of primary object, despite appearance variation and sporadity of detection. The comparison against the existing lowest/highest error rates of video object segmentation methods shows that the object regions generated by efficient graph transduction learning alone can be regarded as coarse segmentation, even without the pixel-based object segmentation described in Sec. 4. The qualitative evaluation of primary object proposals in Fig. 4 further confirms the advantages of the proposed method in SegTrack dataset.

We also compare the object proposals generated from our graph transduction learning with the 'key-segments' generated by Lee *et al.* [9]. Fig. 5 shows the per-frame ground-truth overlap score of those generated object proposals from both methods on SegTrack dataset. The results clearly demonstrate that our method can generate object proposals which are not only temporally dense in each frame, but also break the lower-bound posed by the accuracy of the region candidates produced by [25] by learning a holistic appearance model (note that most of the blue bars are taller than the corresponding red bars in Fig. 5).

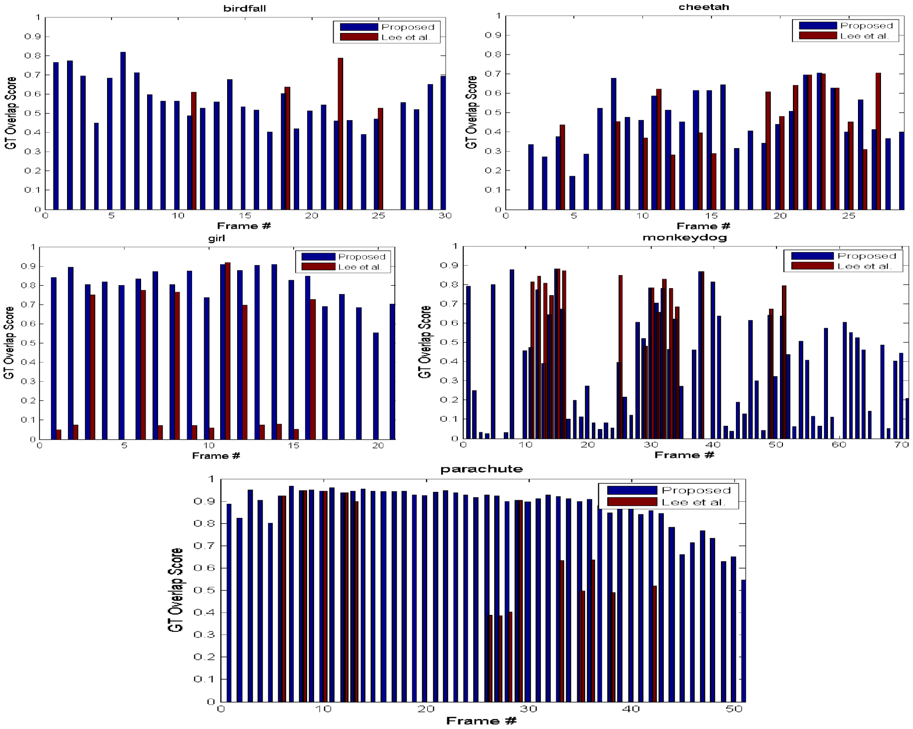

Fig. 5. Ground-truth overlap score of our object proposals and the 'key-segments' from Lee *et al.* [9].

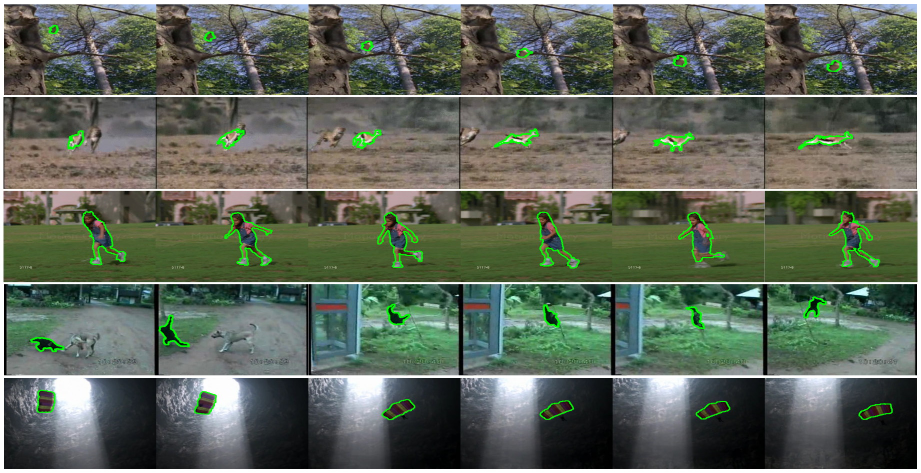

Fig. 6. Segmentation results on SegTrack dataset. The contour of segmented primary object is shown in green.

Table 2. Quantitative segmentation results on SegTrack. Segmentation error as measured by the average number of incorrect pixels per frame. Lower values are better. The best result is shown in red and second best in blue

Video (No. frames)	Ours	[12]	[21]	[11]	[10]	[9]	[4]	[1]
birdfall (30)	**151**	188	217	155	189	288	252	454
cheetah (29)	672	983	890	**633**	806	905	1142	1217
girl (21)	**1121**	1573	3859	1488	1698	1785	1304	1755
monkeydog (71)	359	558	**284**	365	472	521	563	683
parachute (51)	204	339	855	220	221	**201**	235	502
Average	**413**	614	876	452	542	592	594	791
Supervision	N	N	N	N	N	N	Y	Y

Table 3. Quantitative segmentation results on Sports dataset

Video (No. frames)	Ours	Lee *et al.* [9]	Zhang *et al.* [11]
gymnastic (100)	**523**	1595	1951
jump (105)	**364**	1261	3456
monkeybar (200)	**833**	1496	2108
waterski (48)	**1582**	2107	3084
yunakim (200)	**319**	907	4038

Evaluation of Video Object Segmentation We compare our video object segmentation method with five state-of-the-art unsupervised methods [9–11, 21, 12] and two supervised methods [4, 1]. Following [11, 10], we also compute the average number of incorrect pixels over all frames in the five videos as they are roughly of the same frame size. Our method achieves the lowest average number of per-frame pixel error along with superior performance on two out of five videos compared with all 7 state-of-the-art methods with or without supervision. It produces second best results on the rest three videos. Note that our method consistently segments all the videos with low error rate which reflects its robustness on various challenging situations. As a contrast, previous 'object proposal' based methods are limited to the existing region candidates which contain a large amount of label noise.

6.2 Sports Dataset

We have manually generated ground-truth for a new dataset collecting videos from other datasets for video object segmentation. The dataset is challenging: those videos are generally longer than SegTrack dataset; person's varying poses cause frequent self-occlusions and consequently appearance variations; some persons move fast so causing blur whilst some are slow which is very hard to perform motion segmentation. We find that the results on longer and complex videos can better demonstrate the strength of our approach, especially in dealing with fast appearance variation, cluttered scene and complex motions.

We firstly compare the proposed approach with Lee *et al.* [9] which is one of the state-of-the-art 'object proposal' approach, both quantitatively and qualitatively[1]. Table 3 shows the segmentation error on five videos of Sports dataset, comparing our method with [9]. Our method substantially outperforms [9] with low segmentation error across all videos. The qualitative comparison in Fig. 7 further confirms the advantages of the proposed method over [9]. In *gymnastic* (first video), the appearance of the athlete varies quickly due to the fast motion and pose variation. The sparse and noisy 'key-segments' generated by [9] can no longer deal with this complex situation. As a contrast, our approach robustly segments the athlete based on rich descriptions of the primary object regardless of the video length and appearance variation. Similar situations are also present in *monkeybar* (third video), *waterski* (fourth video) and *yunakim* (fifth video) where, in meanwhile, self-occlusion aggravates the failure of [9], due to the lack of prior knowledge in the corresponding frames. The result on *jump* (second video) demonstrates that our method can stably segment small object while preserving temporal coherence (see the missegmentations in the background from [9]).

We also quantitatively and qualitatively compare with Zhang *et al.* [11] on Sports dataset[2]. The quantitative and qualitative comparisons are shown in Table 3 and Fig. 7 respectively. Using local motion-warped overlapping to form new object regions from the region candidates produced by [25], [11] tends to produce either under- or over-segmentations (e.g. the *gymnastic*, *jump* and *yunakim* sequences) due to the spurious object regions and heavy reliance on accurate motion estimation. Zhang *et al.* [11] further assume all object-like regions within each frame are independent and do not explicitly consider spatial affinity, which substantially limits the size of the object region especially when the primary object is comprised of multiple regions with distinct appearances (e.g. the *monkeybar* sequence). Distinctively, our method learns a holistic appearance model to diffuse the prior knowledge from the initial region candidates using graph transduction learning and thus can cope with more complex scenes in natural videos.

7 Conclusion

We have proposed a novel unsupervised video object segmentation method by generating a diverse set of video object proposals in a bottom-up approach. This set of rich descriptions underpin robust segmentations against the large variations of appearance, shape and occlusion in natural videos. The generation of dense video object proposals is cast as performing efficient graph transduction learning based on a holistic appearance model to describe the object-like regions, incorporating both spatial and temporal cues. The proposed approach exhibits superior performance in comparison with the state of the art on the SegTrack dataset and additional challenging data sets posing different challenges.

[1] We used the publicly available source code from:
http://vision.cs.utexas.edu/projects/keysegments/code/

[2] We used the publicly available source code from:
http://dromston.com/projects/video_object_segmentation.php

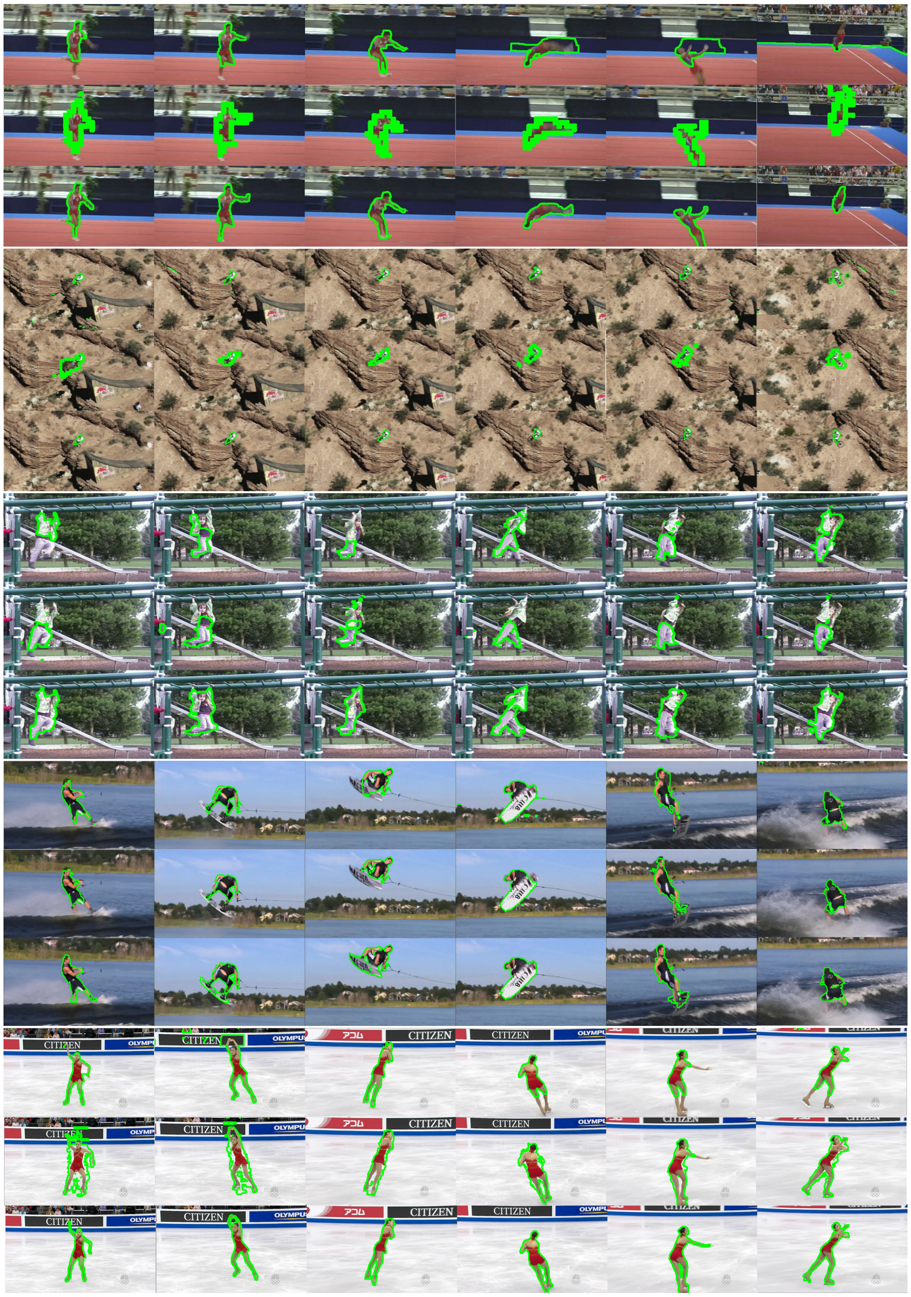

Fig. 7. Segmentation results on Sports dataset. Row 1: Segmentation results by Lee *et al.* [9]. Row 2: Segmentation results by Zhang *et al.* [11]. Row 3: Segmentation by the proposed method.

References

1. Chockalingam, P., Pradeep, S.N., Birchfield, S.: Adaptive fragments-based tracking of non-rigid objects using level sets. In: ICCV. (2009) 1530–1537
2. Bai, X., Wang, J., Simons, D., Sapiro, G.: Video snapcut: robust video object cutout using localized classifiers. ACM Trans. Graph. **28** (2009)
3. Price, B.L., Morse, B.S., Cohen, S.: Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In: ICCV. (2009) 779–786
4. Tsai, D., Flagg, M., Nakazawa, A., Rehg, J.M.: Motion coherent tracking using multi-label mrf optimization. International Journal of Computer Vision **100** (2012) 190–202
5. Wang, T., Collomosse, J.P.: Probabilistic motion diffusion of labeling priors for coherent video segmentation. IEEE Transactions on Multimedia **14** (2012) 389–400
6. Wang, T., Han, B., Collomosse, J.P.: Touchcut: Fast image and video segmentation using single-touch interaction. Computer Vision and Image Understanding **120** (2014) 14–30
7. Sheikh, Y., Javed, O., Kanade, T.: Background subtraction for freely moving cameras. In: ICCV. (2009) 1219–1225
8. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV (5). (2010) 282–295
9. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV. (2011) 1995–2002
10. Ma, T., Latecki, L.J.: Maximum weight cliques with mutex constraints for video object segmentation. In: CVPR. (2012) 670–677
11. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: CVPR. (2013) 628–635
12. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video Segmentation by Tracking Many Figure-Ground Segments. In: ICCV. (2013)
13. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Sch, B.: Learning with local and global consistency. In: NIPS. Volume 1. (2004)
14. Wang, J., Xu, Y., Shum, H.Y., Cohen, M.F.: Video tooning. ACM Trans. Graph. **23** (2004) 574–583
15. Collomosse, J.P., Rowntree, D., Hall, P.M.: Stroke surfaces: Temporally coherent artistic animations from video. IEEE Trans. Vis. Comput. Graph. **11** (2005) 540–549
16. Brendel, W., Todorovic, S.: Video object segmentation by tracking regions. In: ICCV. (2009) 833–840
17. Huang, Y., Liu, Q., Metaxas, D.N.: Video object segmentation by hypergraph cut. In: CVPR. (2009) 1738–1745
18. Reina, A.V., Avidan, S., Pfister, H., Miller, E.L.: Multiple hypothesis video segmentation from superpixel flows. In: ECCV (5). (2010) 268–281
19. Grundmann, M., Kwatra, V., Han, M., Essa, I.A.: Efficient hierarchical graph-based video segmentation. In: CVPR. (2010) 2141–2148
20. Xu, C., Xiong, C., Corso, J.J.: Streaming hierarchical video segmentation. In: ECCV (6). (2012) 626–639
21. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV. (2013)
22. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012)

23. Greenspan, H., Goldberger, J., Mayer, A.: A probabilistic framework for spatio-temporal video representation & indexing. In: ECCV. (2002) 461–475
24. Wang, J., Thiesson, B., Xu, Y., Cohen, M.F.: Image and video segmentation by anisotropic kernel mean shift. In: ECCV (2). (2004) 238–249
25. Endres, I., Hoiem, D.: Category independent object proposals. In: ECCV. (2010) 575–588
26. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: ECCV. (2004) 25–36
27. Jebara, T., Wang, J., Chang, S.F.: Graph construction and *b*-matching for semi-supervised learning. In: ICML. (2009) 56
28. Dollar, P., Zitnick, C.L.: Structured forests for fast edge detection. In: ICCV. (2013)
29. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23** (2001) 1222–1239
30. VOT2013: The vot2013 challenge dataset. http://www.votchallenge.net (2013)