Computer Vision and Image Understanding 143 (2016) 159–172

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Primary object discovery and segmentation in videos via graph-based transductive inference

Huiling Wang [a], Tinghuai Wang [b,*]

[a] Lappeenranta University of Technology, Lappeenranta 53850, Finland
[b] Nokia Technologies, Visiokatu 3, Tampere 33720, Finland

ARTICLE INFO

Article history:
Received 2 October 2014
Accepted 15 November 2015

Keywords:
Graph-based transductive inference
Video object segmentation
Object proposal

ABSTRACT

The proliferation of video data makes it imperative to develop automatic approaches that semantically analyze and summarize the ever-growing massive visual data. As opposed to existing approaches built on still images, we propose an algorithm that detects recurring primary object and learns cohort object proposals over space-time in video. Our core contribution is a graph transduction process that exploits both appearance cues learned from rudimentary detections of object-like regions, and the intrinsic structures within video data. By exploiting the fact that rudimentary detections of recurring objects in video, despite appearance variation and sporadity of detection, collectively describe the primary object, we are able to learn a holistic model given a small set of object-like regions. This prior knowledge of the recurring primary object can be propagated 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. We present extensive experiments on challenging datasets that demonstrate the superior performance of our approach compared with the state-of-the-art methods.

1. Introduction

Segmenting object from video remains an open challenge with recent advances relying upon prior knowledge supplied via interactive initialization or correction [1–6]. Yet fully automatic video object discovery and segmentation [7–12] remains useful in scenarios where the human in the loop is impractical, such as video summarization or ingest pre-processing for video indexing or recognition. 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 primary video object discovery and 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*. Following [9–14], the primary video object refers to the object that presents saliently, in terms of either appearance or motion, in most of the frames. 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 [9–13].

Our strategy is to create feature-based rudimentary detections of regions for the primary object by learning from weakly labelled examples of object-like regions. These detections serve as informative indicators of the appearance and location of the object. We propagate this learned prior knowledge on an undirected space-time graph consisting of regions, solving the transduction learning efficiently with a fast convergence technique [15]. Inference at the region level further makes our dense video object proposal extraction approach a practical solution for automatic object segmentation on natural video sequences.

We describe our proposed video object proposal algorithm in Section 3, presenting the utilization of video object proposals for robust video object segmentation in Section 4. In Section 6, we evaluate our video object proposal and segmentation approach on benchmark dataset and additional dataset comprising challenging video clips exhibiting clutter, occlusion and agile motion, comparing against state-of-the-art semi-automatic and automatic algorithms.

* Corresponding author.
E-mail addresses: huiling.wang@lut.fi (H. Wang), tinghuai.wang@gmail.com, tinghuai.wang@nokia.com (T. Wang).

http://dx.doi.org/10.1016/j.cviu.2015.11.006
1077-3142/

2. Related work

Generic object detection has been intensively studied in context of still images recently [16–21]. Alexe et al. [16] introduced the objectness measure which computes the probability that a window contains any object, using a Bayesian classifier based on multiple cues. Carreira et al. [17] (CPMC) proposed to use several graphcuts running using random positive and negative seeds. Each generated foreground mask serves as an object proposal, and the proposals are ranked according to a learned scoring function. Similarly to CPMC, Endres and Hoiem [18] proposed to generate multiple foreground segmentations and use these as object proposals using binary CRF segmentation with random seeds. [19] performs an ad-hoc hierarchical bottom-up agglomeration of groups of regions and a fixed number of proposals are generated at each step of the agglomeration. Manen et al. [20] proposed an approach based on randomly growing groups of regions, which allows to generate any desirable number of object proposals. Cheng et al. [21] introduced a simple and fast objectness measure to compute the objectness of each image window at various scale and aspect ratio. The bounding box based objectness measure methods [16,21] share the similar limitations that a bounding box might not localize the object instances as accurately as a segmentation region. Generating object proposals incorporating temporal information has been receiving more attentions recently [22,23]. Sharir and Tuytelaars [22] proposed to extract object proposals in each frame separately which are linked across frames into object hypotheses. This approach suffers from the mis-segmentations of object proposals in each independent frame. Oneata et al. [23] proposed a supervoxel method for spatio-temporal detection. However, supervoxel based approaches usually become computationally infeasible for pixel counts in even moderate size videos, and often undersegment small or fast moving objects that form disconnected space-time volumes.

Our method follows the segmentation based approach to generating video object proposals which provides a set of rich descriptions underpinning robust segmentation and many other applications against large variations of appearance, shape and occlusion in natural videos. As apposed to those image based generic object detection algorithms which typically generate an excessive amount of proposals ($> 10^4$), our approach generates cohort object proposals over space-time to capture the essential parts of tentative primary object exploring cues beyond the single still image.

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 [1–6,24] or perform spatio-temporal grouping [25,26]. 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 video object segmentation 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 [12,27–31] take a bottom-up approach based on spatio-temporal appearance and motion constraints. Motion segmentation methods [32–38] cluster pixels or regions 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. Occlusion boundary approach [37] has been proposed to detecting occlusion boundaries and assigning figure/ground labels to both sides of those boundaries. Layered models have been studied in [32,35,36,39,40]. Methods which generate over-segmentations for later processing analog to still-image regions [41] have also been proposed [42,43], by applying spatio-temporal clustering based on low level features. Papazoglou and Ferrari [12] determine an initial set of foreground pixels based purely on motion and refine the FG/BG labels using Graph Cut. 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 [9–11,13,14] are proposed based on exploring recurring object-like regions from still images by measuring generic object appearance [18]. 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 [18] 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. [13] 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. Yang et al. [14] proposed a method to fuse appearance and motion saliency maps for discovering primary video object. All the above methods do not build an explicit holistic appearance model but relies on local heuristics and motion for selecting and merging the object proposals or saliency maps.

To address the limitations of the above approaches [9–11,13,14], 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 segmentation is driven by Markov Random Field (MRF) approach. A variety of MRF models as well as inference and learning methods have been developed for addressing numerous computer vision problems during the past decade. MRF models can be categorized into pairwise models and higher-order models. Various works have investigated the modeling of vision problems using pairwise MRFs (e.g., [36,44–48]) and the efficient inference in pairwise MRFs

(e.g., [49–53]). Higher-order MRFs has recently increased largely the ability of graph-based modeling and better characterized the statistics between random variables [54–57]. For more complete surveys of Markov Random Field, the reader is refereed to [58,59].

To summarize our contributions: (1) we introduce a graph transduction learning approach to efficiently detect recurring objects and learn cohort object proposals over space-time in video, by exploring the holistic patterns of primary object collectively revealed by a small set of object-like regions and the intrinsic structure within video data (2) we utilize this set of object proposals which provide sufficient and diverse appearance, shape, and location prior information to drive the object segmentation problem while preserving spatio-temporal coherence.

3. Video object proposals

We formulate the problem of generating a diverse set of dense video object proposals as a graph transduction learning problem, given the prior knowledge from the initial detection and the intrinsic structure within data. The purpose of graph transduction is to propagate the sparse local evidence over space-time in video to coherently locate strong valid indicators of primary object.

3.1. Graph transduction learning of object proposals

Let $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^m$ denote a given dataset, and $\mathcal{L}$ be a continuous label set. We assume that a subset of the dataset ($\bar{\mathcal{X}} \subseteq \mathcal{X}$) have been noisily labeled by values in $\mathcal{L}$, and the remaining data points $\mathcal{X} \backslash \bar{\mathcal{X}}$ might be unlabeled. Our goal of transduction learning is to predict the label of both labeled and unlabeled points given the initial noisy label.

3.1.1. Space-time graph of regions

To perform transduction learning, we define a weighted space-time graph $\mathcal{G}_s = (\mathcal{V}, \mathcal{E})$ spanning $\mathcal{X}$, i.e. 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 [60] 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

$$W_{ij} = \exp\left(-\frac{\chi^2(h_{r_i}, h_{r_j})}{2\beta}\right),$$

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 [61], we set all W_{ij} are set to zero if r_i and r_j are not adjacent.

3.1.2. Graph transduction learning

Graph transduction learning propagates label information from labeled nodes to unlabeled nodes. Let the node degree matrix $D = \text{diag}([d_1, \ldots, d_N])$ be defined as $D_i = \sum_{j=1}^{N} W_{ij}$, where $N = |\mathcal{V}|$. We follow a similar formulation with [15] to minimize an energy function $E(F)$ with respect to all region labels F:

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

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 [15]. 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

$$(I - (1-\gamma)S)F^* = \gamma Y. \tag{4}$$

An optimal solution for F can be solved using the Conjugate Gradient method with very fast convergence.

We use the predictions from the holistic object model (described in Section 3.2) 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:

$$Y_+ = \begin{cases} Y & \text{if } Y > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and

$$Y_- = \begin{cases} -Y & \text{if } Y < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Combining the diffusion processes of both the object-like regions and background can produce more efficient and coherent labelling, taking advantage of their complementary properties. We perform the optimization for two diffusion processes simultaneously as follows:

$$F^* = \gamma (I - (1-\gamma)S)^{-1} (Y_+ - Y_-). \tag{7}$$

This enables a faster and stable optimization avoiding separate optimizations. 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. 1.

3.2. Learning a holistic appearance model

We describe the process of learning a holistic appearance model of the primary object for graph transduction learning in this section (Fig. 2). The process begins by discovering an initial set of object-like regions from video sequence. Throughout the discovery process, we maintain two disjoint sets of image regions: $\mathcal{H}$ and $\mathcal{U}$, which represent the discovered object-like regions and those remain in the general unlabeled pool, respectively. $\mathcal{H}$ is initially empty whilst $\mathcal{U}$ is set to be all the 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 [18]

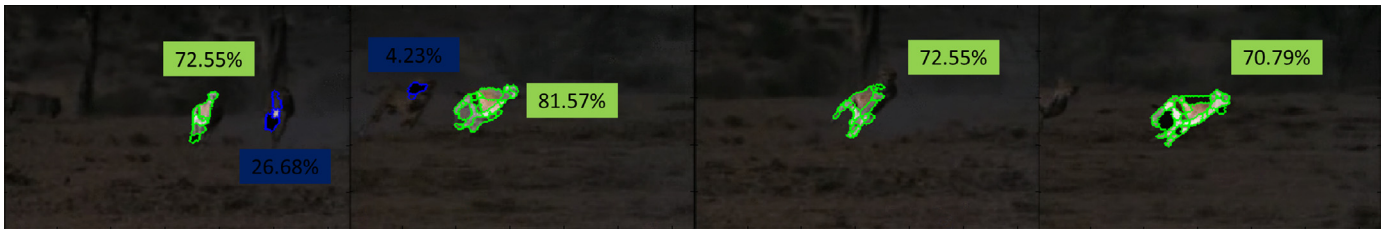

Fig. 1. Exemplar video object proposals from CHEETAH sequence. Colors of contour indicate different proposals. The transparency of each region indicates the objectness (F) from graph transduction learning. The objectness of each final object proposal is computed by averaging the constituent region-wise objectness F weighted by area.

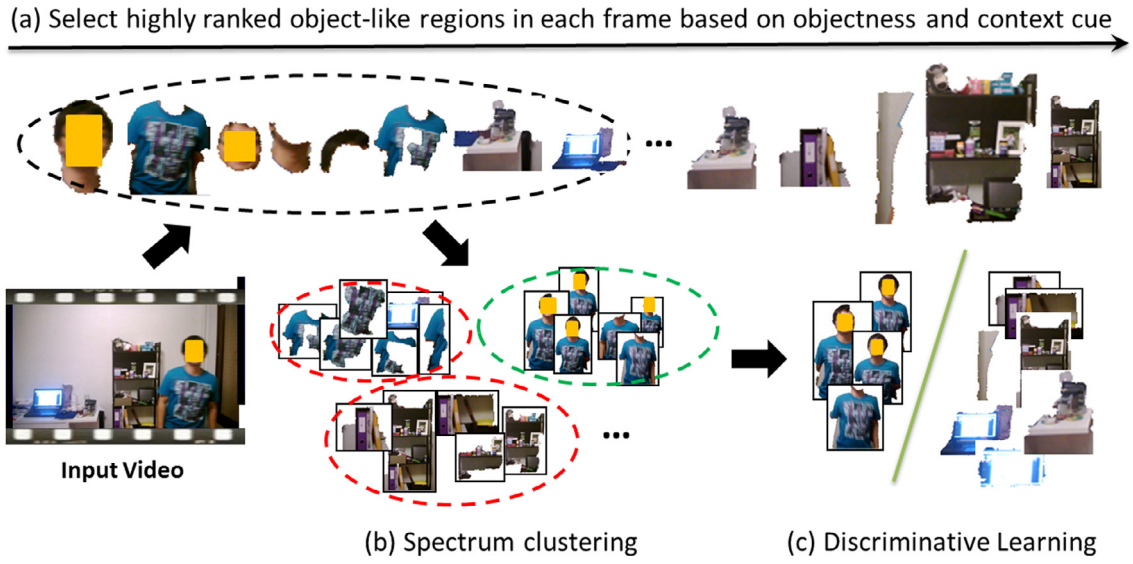

Fig. 2. The process of learning a holistic appearance model of the primary object.

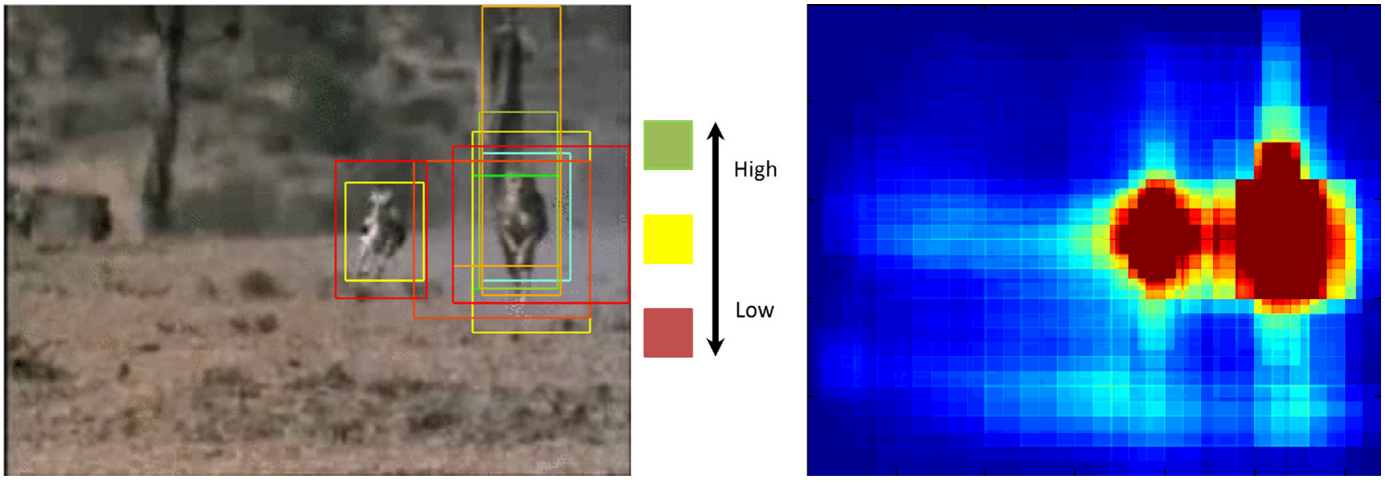

Fig. 3. Computing the saliency map: (a) top-scoring detection windows using [16]; (b) saliency map by accumulating 1000 windows.

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 a score $\Phi(r)$ for each region r as

$$\Phi(r) = Obj(r) + Ctx(r) \tag{8}$$

where $Obj(r)$ is the objectness cue and $Ctx(r)$ is the context cue.

3.2.1. Objectness cue

The static intra-frame objectness score $Obj(r)$ quantifies how likely it is for an image region or window to contain an object of any class. Note, this objectness score does *not* consider context cues, e.g. motion, object categories and temporal coherence etc., and reflects only the generic object-like properties of the region (saliency, apparent separation from background, etc.). $Obj(r)$ is defined as:

$$Obj(r) = \mathcal{A}(r) + \mathcal{S}(r)$$

where $\mathcal{A}(r)$ indicates region level appearance score computed using [18] and $\mathcal{S}(r)$ represents the saliency level of region r which is defined as:

$$\mathcal{S}(r) = \theta_r \sum_{x \in r} S(x).$$

S denotes a real valued saliency map the same size as the input image, and θ_r is an adaptive weight indicating the 'purity' of salience

Fig. 4. 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; (d) per-pixel object probabilities from GMM color model trained using object proposals.

inside region r, which is computed using the variance of saliency of all constituent pixels of r:

$$\theta_r = \exp\left(-\frac{1}{|r|}\sum_{x\in r}(S(x) - \frac{1}{|r|}\sum_{x\in r}S(x))^2\right).$$

Lower 'purity' or higher variance of salience results in lower region level salience.

To compute the saliency map, we adopt [16] which combines several image cues measuring distinctive characteristics of objects. For each image, we sample 1000 windows likely to contain an object from this measure and set the saliency map S to be the pixel-wise mean of objectness scores of all detected windows.

3.2.2. Context cue

We consider motion as the major context cue to resolve the visual ambiguities present in the objectness cue at this stage. Context cue $Ctx(r)$ reflects the disparity of motions between primary object and background. We compute optical flow [60] histograms for region r and $\bar{r}$ which is formed by merging all the closest surrounding regions of r. We find that using surrounding regions is more informative than using pixels in a loosely fit bounding box around r. We compute $Ctx(r)$ as

$$Ctx(r) = 1 - \exp(-\chi^2_{flow}(r, \bar{r})),$$

where $\chi^2_{flow}(r, \bar{r})$ is the χ^2 distance between L_1-normalized optical flow histograms for regions r and $\bar{r}$.

3.2.3. Discovery of object-like regions

A candidate pool $\mathcal{C}$ can be formed by taking the top K highest-scoring regions from each frame, and then identify groups of object-like regions that may represent a foreground object by performing spectral clustering [62] in $\mathcal{C}$. All clusters are ranked based on the average score Φ (Eq. (8)) 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 are added to $\mathcal{H}$.

Each object-like region 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, i.e. the SVM margin. All weights are normalized between -1 and 1, by the sum of positive and negative margins.

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 our proposed graph transduction learning approach (Section 3.1), exploiting both the intrinsic structure within data and the initial local evidence from the holistic appearance model. Fig. 4(a–c) 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.

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:

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

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 [63], $||c_i - c_j||^2$ 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:

$$\psi_i(x_i) = -\log(w\cdot U^c_i(x_i) + (1 - w)\cdot U^o_i(x_i))$$

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

4.1. 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

Fig. 5. Five sequences used for setting parameters.

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. Fig. 4(d) shows per-pixel object probabilities from GMM color model.

4.2. 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 related to the objectness value (F in (7)) of the region it belongs to:

$$U_i^o(x_i) = \begin{cases} F_i & \text{if } x_i = 1 \\ 1 - F_i & \text{if } x_i = 0 \end{cases} \tag{9}$$

5. Implementation details

We start by computing feature descriptors for all the regions in video. We utilize the superpixel regions returned from [18] which is produced by [64]. We select the top $K = 10$ highest-scoring regions from each frame to form a candidate pool $\mathcal{C}$.

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 set of 5 videos from VOT2013 [65] and VBR [66] shown in Fig. 5 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 three datasets[1]: SegTrack [4], Sports (a new dataset consisting of five videos), and YouTube-Objects [67]. Two videos (*waterski, yunakim*) of this new dataset are from GaTech video segmentation dataset [30], two (*jump, gymnastic*) from the challenging VOT2013 [65] dataset, and one (*monkeybar*) from video tooning [25].

[1] Results can be viewed online at: http://youtu.be/mrwl8l6t0HU.

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 on SegTrack and Sports as the average percentage of per-frame pixel error compared to the ground-truth:

$$\text{error} = \frac{\text{XOR(S,GT)}}{\text{NF} \cdot \text{P}} \tag{10}$$

where S denotes the label for every pixel in the video, GT is the ground-truth, NF is the total number of frames in the video, and P is the total number of pixels per frame. YouTube-Objects dataset provides ground-truth bounding boxes on the object of interest in one frame for each of 1407 video shots. We adopt the performance measure used in [12,67], i.e. the bounding box intersection-over-union ratio. We automatically fit a bounding box to the largest connected component in the segmentation output by our method for the purpose of this evaluation.

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 [9–13] 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.

6.1.1. Ablation studies

To understand the contribution of each proposed modules in our algorithm, we compare the segmentation results using the proposed method against two baseline schemes (I) unary term using only color likelihood from GMM color models trained on object proposals and (II) unary term using only objectness value from Eq. (9). The quantitative results on SegTrack dataset are listed in Table 1, where the column *GMM* refers to the results from scheme I and *Objectness* refers to scheme II respectively. We observe that the segmentation results driven by objectness value are only slightly inferior to the results using the full system, which demonstrates the efficacy of the generated object proposals. Although GMM based unary underperforms objectness based unary, its complementary effect on objectness-only unary can be clearly observed by comparing the full-system and objectness-only results. More specifically, objectness gives more accurate yet occasionally unsmooth predictions, as shown in Fig. 4 (i.e., the girl's hair); GMM color model gives more coherent predictions to compensate unsmooth objectness values whereas it suffers from similar colors. This noise in GMM unary is properly tackled by the pairwise term in graph cut optimization.

To evaluate our method's capability to detect and generate spatio-temporal coherent and dense object-like regions, we further compare the generated proposals with [18], one of the state-of-the-art segment based object proposal methods on still images, as the baseline. Table 1 also compares the per-pixel error rate of our object proposals (column *Proposals*), per-frame best scoring object proposal generated from [18], and also the lowest/highest error rates of all existing methods on SegTrack dataset. We observe that [18] returns inconsistent

Fig. 6. Primary object proposals in SegTrack dataset. Row 1: top-scoring object proposal by Endres and Hoiem [18] in each frame. Row 2: primary object proposal generated by the proposed graph transduction learning method.

and sporadic object 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 Section 4. The qualitative comparison in Fig. 6 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. 7 shows the per-frame ground-truth overlap score of those

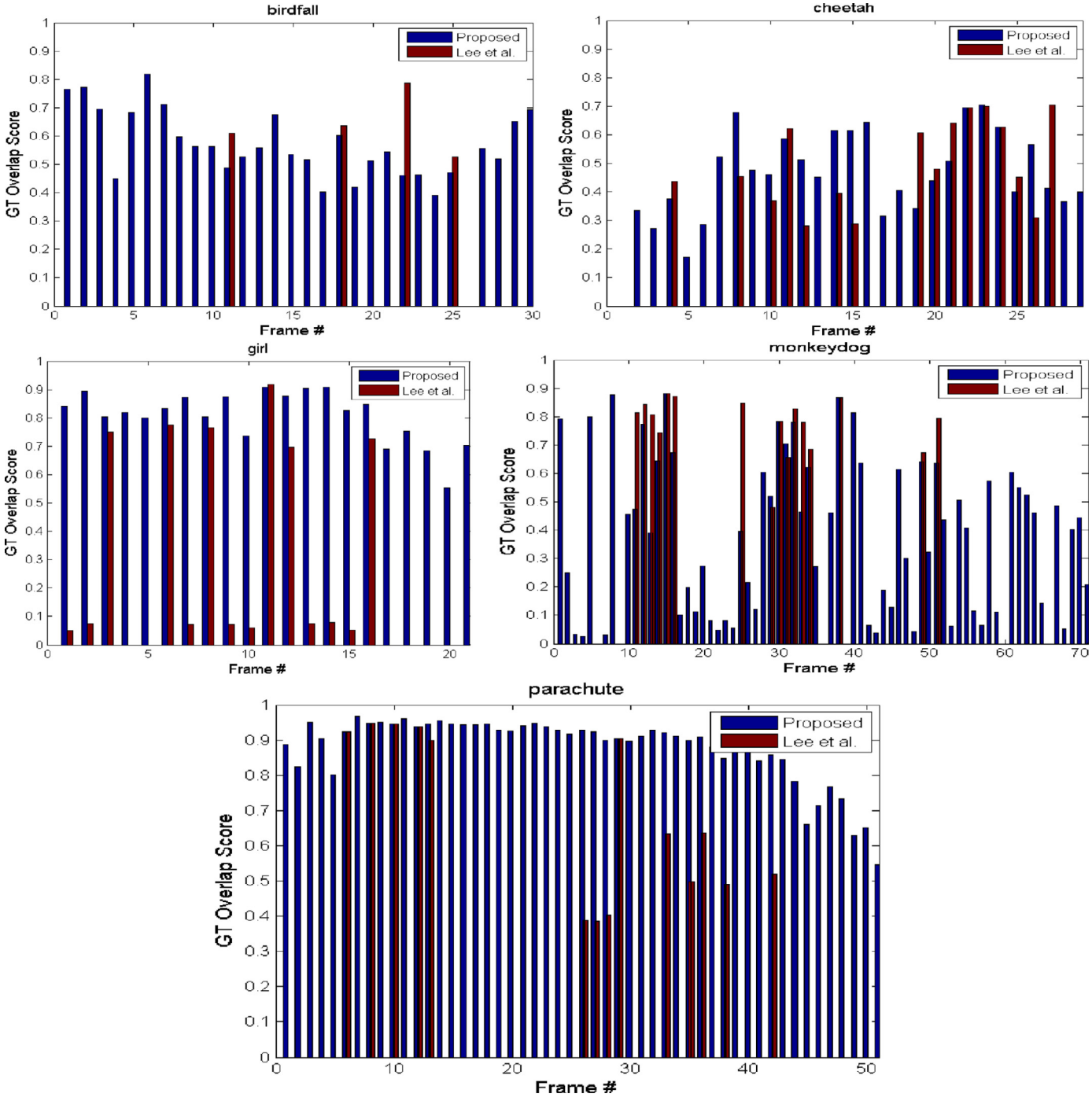

Fig. 7. Ground-truth overlap score of our object proposals and the 'key-segments' from Lee et al. [9].

Table 1
Ablation studies on SegTrack. Segmentation results using the proposed algorithm (column *Full System*), GMM color model based unary term only (column *GMM*) and objectness value based unary only (column *Objectness*) are compared. The proposed video object proposals are also compared with the per-frame top-scoring object proposal from [18], and also the lowest/highest error rates of all existing video object segmentation methods. Segmentation error as measured by the average percentage (%) of incorrect pixels per frame.

Video (no. of frames)	Full system	GMM	Objectness	Proposals	[18]	Lowest	Highest
Birdfall (30)	0.17	0.42	0.21	0.30	26.17	0.17	0.55
Cheetah (29)	0.86	1.25	0.92	1.10	26.89	0.51	2.56
Girl (21)	0.85	1.55	1.07	1.28	6.39	0.85	5.93
Monkeydog (71)	0.44	1.52	0.58	1.08	37.84	0.37	1.87
Parachute (51)	0.14	0.68	0.35	0.30	56.91	0.13	1.09

Table 2
Quantitative segmentation results on SegTrack. Segmentation error as measured by the average percentage (%) of incorrect pixels per frame. Lower values are better.

Video	Ours	[24]	[13]	[12]	[11]	[10]	[68]	[9]	[8]	[4]	[1]
Birdfall	**0.17**	0.29	0.22	0.26	0.18	0.22	0.55	0.34	0.55	0.30	0.54
Cheetah	0.86	**0.51**	1.28	1.16	0.82	1.05	1.53	1.18	2.56	1.49	1.58
Girl	**0.85**	1.51	1.23	3.01	1.16	1.33	4.44	1.39	5.93	1.02	1.37
Monkeydog	0.44	0.65	0.73	**0.37**	0.48	0.61	1.87	0.68	1.87	0.73	0.89
Parachute	0.14	**0.13**	0.23	0.59	0.15	0.15	1.09	0.14	0.76	0.16	0.34
Average	**0.42**	0.53	0.65	0.79	0.47	0.52	1.70	0.61	1.92	0.64	0.84
Supervision	N	Y	N	N	N	N	N	N	N	Y	Y

Fig. 8. Segmentation results on SegTrack dataset. The contour of segmented primary object is shown in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 [18] by learning a holistic appearance model (note that most of the blue bars are taller than the corresponding red bars in Fig. 7).

6.1.2. Evaluation of video object segmentation

We compare our video object segmentation method with five state-of-the-art automatic methods [9–13], three semi-automatic methods [1,4,24] and two motion segmentation methods [8,68]. Our method achieves the lowest average percentage of per-frame pixel error along with superior performance on two out of five videos compared with all 10 state-of-the-art video object segmentation methods with or without supervision and motion segmentation methods. It produces second best results on two out of 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.

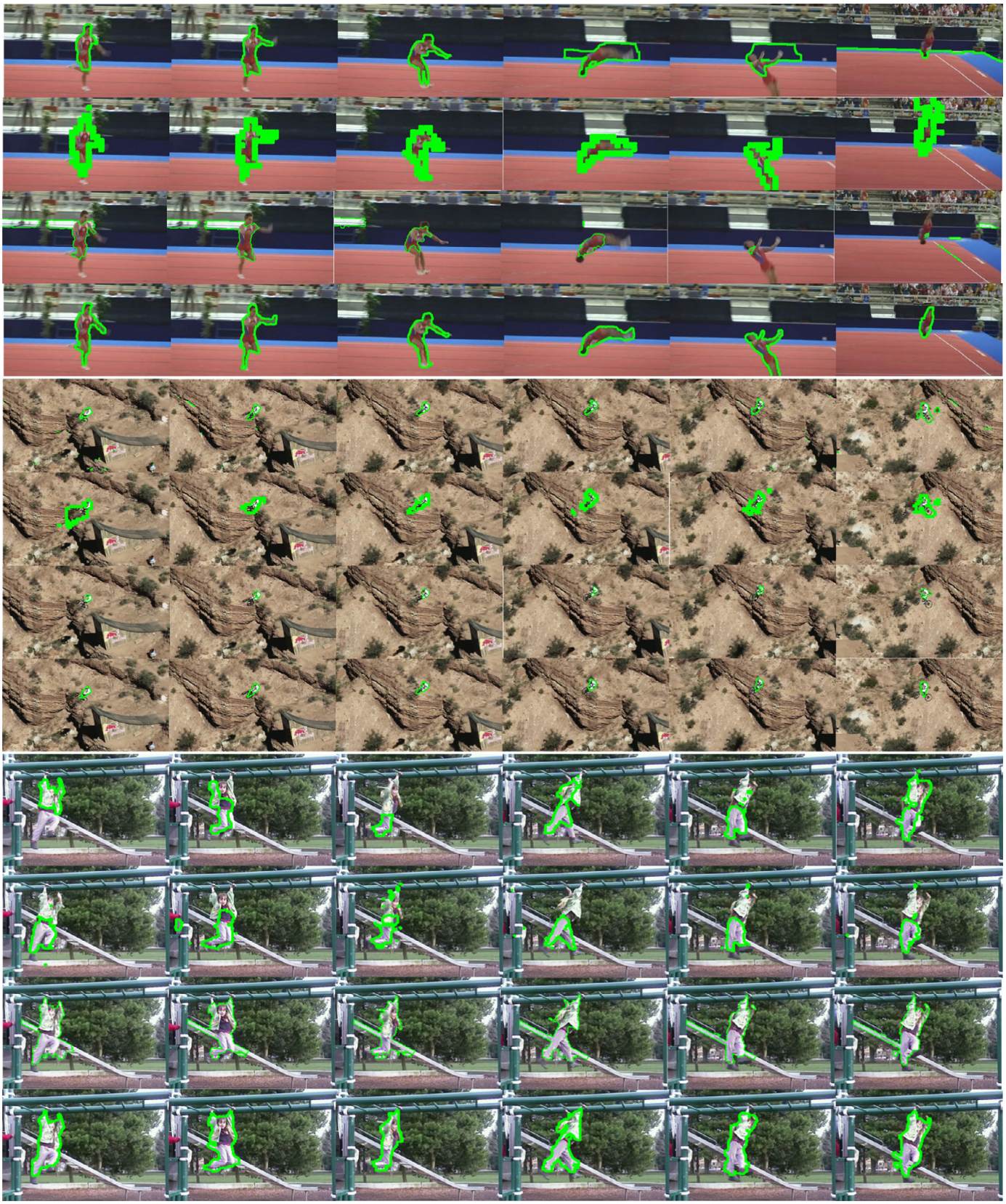

Fig. 9. Segmentation results on *gymnastic*, *jump* and *monkeybar* from Sports dataset. Row 1: Segmentation results by Lee et al. [9]. Row 2: segmentation results by Zhang et al. [11]. Row 3: segmentation results by Papazoglou and Ferrari [12]. Row 4: segmentation by the proposed method.

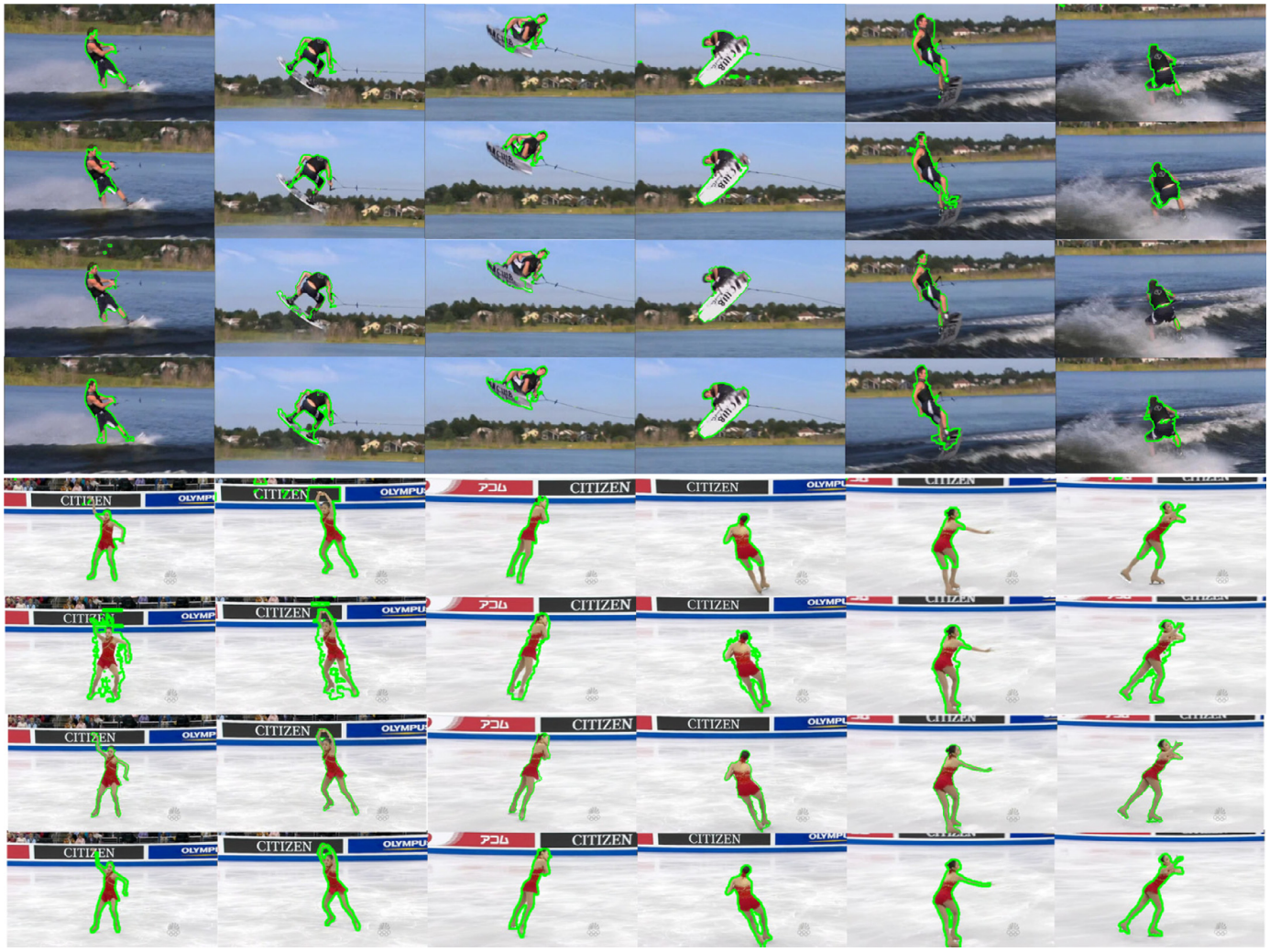

Fig. 10. Segmentation results on *waterski* and *yunakim* Sports dataset. Row 1: segmentation results by Lee et al. [9]. Row 2: segmentation results by Zhang et al. [11]. Row 3: segmentation results by Papazoglou and Ferrari [12]. Row 4: segmentation by the proposed method.

Table 3
Quantitative results on Sports dataset. Segmentation error as measured by the average percentage (%) of incorrect pixels per frame.

Video (no. of frames)	Ours	[9]	[11]	[12]
Gymnastic (100)	**0.88**	2.77	3.39	4.73
Jump (105)	**0.15**	0.55	1.5	0.44
Monkeybar (200)	**1.65**	3.03	4.27	2.65
Waterski (48)	**0.98**	1.38	2.02	2.79
Yunakim (200)	**0.35**	1.06	4.72	0.45

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.[2] 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 comparisons in Figs. 9 and 10 further confirm the advantages of the proposed method over [9]. In *gymnastic* (Fig. 9 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* (Fig. 9 third video), *waterski* (Fig. 10 first video) and *yunakim* (Fig. 10 second 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* (Fig. 9 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.[3] The quantitative and qualitative comparisons are shown in Table 3 and Fig. 9, respectively. Using local motion-warped overlapping to form new object regions from the region candidates produced by [11,18] 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

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

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

Fig. 11. Example results for 10 categories from YouTube-Objects dataset.

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.

Our method also outperforms Papazoglou and Ferrari [12][4] which is one of the state-of-the-art approaches utilizing 'occlusion boundaries'. Heavily relying on motion estimations, Papazoglou and Ferrari [12] is sensitive to local erroneous optical flows caused by similar colors (e.g. the *monkeybar* sequence) or strong motion blurs (e.g. the *gymnastic* sequence). In contrast, our method handles these situations better with a higher level notion of object, enabled by the diverse set of dense video object proposals.

6.3. YouTube-objects dataset

YouTube-Objects [67] is a large-scale dataset which consists of 126 completely unconstrained and very challenging videos from 10 object classes. The videos feature large camera motion, diverse backgrounds, illumination changes; the objects undergo rapid movement, strong scale and viewpoint changes, non-rigid deformations.

We compare to [8,12,23,67], and report their performance as originally stated in [12,67]. As shown in Table 4, our method outperforms

[4] We used the publicly available source code from: http://calvin.inf.ed.ac.uk/software/fast-video-segmentation.

Table 4
Intersection-over-union overlap accuracies on YouTube-Objects Dataset.

Method	Plane	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Avg
Ours	0.63	**0.69**	**0.40**	0.61	**0.48**	**0.46**	**0.67**	**0.53**	**0.47**	**0.38**	**0.53**
[8]	0.54	0.20	0.38	0.37	0.32	0.29	0.27	0.35	0.45	0.38	0.35
[67]	0.52	0.18	0.34	0.35	0.22	0.18	0.14	0.27	0.41	0.25	0.29
[12]	**0.65**	0.67	0.39	**0.65**	0.46	0.40	0.65	0.48	0.39	0.25	0.50

the competing methods in 8 out of 10 classes, with gains up to 0.033 in average accuracy over the best competing method [12], which confirms what we observed on the SegTrack and Sports datasets. We also compare with the video object proposal method presented by Oneata et al. [23] which reported the average accuracies using different number of proposals. The accuracy reported is 0.461 using 10 proposals, a result which is considerably lower than our method (0.534).

7. Discussions

Although the proposed approach can deliver superior performance comparing with existing methods focusing on the primary video object, the system as a whole cannot deal with multiple objects simultaneously due to the non-discriminative nature of generic object detection and large variation of video objects (with occlusions). Consistently discovering, tracking and segmenting *multiple generic objects* in natural videos remains an open question if no prior knowledge is present. Yet it can be relaxed into a category-specific object detection and segmentation problem, given much stronger supervision and prior knowledge [69,70], i.e., the category-dependent object detector and video level label are available.

While further discussion of multiple category-specific objects problem is beyond the scope of this paper, it is worth noting that our core contribution, i.e., graph transduction learning approach for generating object proposals (Section 3.1), is able to generalize to any video object segmentation problem, given proper object detection and holistic modeling as in Section 3.2.

8. Conclusion

We have proposed a novel automatic 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 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, YouTube-Objects dataset, and additional challenging dataset posing different challenges.

References

[1] P. Chockalingam, S.N. Pradeep, S. Birchfield, Adaptive fragments-based tracking of non-rigid objects using level sets, in: ICCV, 2009, pp. 1530–1537.
[2] X. Bai, J. Wang, D. Simons, G. Sapiro, Video snapcut: robust video object cutout using localized classifiers, ACM Trans. Graph. 28 (3) (2009).
[3] B.L. Price, B.S. Morse, S. Cohen, Livecut: learning-based interactive video segmentation by evaluation of multiple propagated cues, in: ICCV, 2009, pp. 779–786.
[4] D. Tsai, M. Flagg, A. Nakazawa, J.M. Rehg, Motion coherent tracking using multi-label MRF optimization, Int. J. Comput. Vis. 100 (2) (2012) 190–202.
[5] T. Wang, J.P. Collomosse, Probabilistic motion diffusion of labeling priors for coherent video segmentation, IEEE Trans. Multimed. 14 (2) (2012) 389–400.
[6] T. Wang, B. Han, J.P. Collomosse, Touchcut: fast image and video segmentation using single-touch interaction, Comput. Vis. Image Underst. 120 (2014) 14–30.
[7] Y. Sheikh, O. Javed, T. Kanade, Background subtraction for freely moving cameras, in: ICCV, 2009, pp. 1219–1225.
[8] T. Brox, J. Malik, Object segmentation by long term analysis of point trajectories, in: ECCV, 2010, pp. 282–295.
[9] Y.J. Lee, J. Kim, K. Grauman, Key-segments for video object segmentation, in: ICCV, 2011, pp. 1995–2002.
[10] T. Ma, L.J. Latecki, Maximum weight cliques with mutex constraints for video object segmentation, in: CVPR, 2012, pp. 670–677.
[11] D. Zhang, O. Javed, M. Shah, Video object segmentation through spatially accurate and temporally dense extraction of primary object regions, in: CVPR, 2013, pp. 628–635.
[12] A. Papazoglou, V. Ferrari, Fast object segmentation in unconstrained video, in: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013, 2013, pp. 1777–1784.
[13] F. Li, T. Kim, A. Humayun, D. Tsai, J.M. Rehg, Video segmentation by tracking many figure-ground segments, in: ICCV, Australia, December 1–8, 2013, 2013, pp. 2192–2199.
[14] J. Yang, G. Zhao, J. Yuan, X. Shen, Z. Lin, B. Price, J. Brandt, Discovering primary objects in videos by saliency fusion and iterative appearance estimation, IEEE Trans. Circuits Syst. Video Technol. (2015).
[15] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Sch, Learning with local and global consistency, in: NIPS, 2004, pp. 321–328.
[16] B. Alexe, T. Deselaers, V. Ferrari, What is an object? in: CVPR, 2010, pp. 73–80.
[17] J. Carreira, C. Sminchisescu, Constrained parametric min-cuts for automatic object segmentation, in: CVPR, 2010, pp. 3241–3248.
[18] I. Endres, D. Hoiem, Category independent object proposals, in: ECCV, 2010, pp. 575–588.
[19] K.E.A. van de Sande, J.R.R. Uijlings, T. Gevers, A.W.M. Smeulders, Segmentation as selective search for object recognition, in: ICCV, 2011, pp. 1879–1886.
[20] S. Manen, M. Guillaumin, L.J.V. Gool, Prime object proposals with randomized prim's algorithm, in: ICCV, 2013, pp. 2536–2543.
[21] M. Cheng, Z. Zhang, W. Lin, P.H.S. Torr, BING: binarized normed gradients for objectness estimation at 300fps, in: CVPR, 2014, pp. 3286–3293.
[22] G. Sharir, T. Tuytelaars, Video object proposals, in: CVPR Workshops, 2012, pp. 9–14.
[23] D. Oneata, J. Revaud, J.J. Verbeek, C. Schmid, Spatio-temporal object detection proposals, in: ECCV, 2014, pp. 737–752.
[24] D. Varas, F. Marques, Region-based particle filter for video object segmentation, in: CVPR, 2014, pp. 3470–3477.
[25] J. Wang, Y. Xu, H.-Y. Shum, M.F. Cohen, Video tooning, ACM Trans. Graph. 23 (3) (2004) 574–583.
[26] J.P. Collomosse, D. Rowntree, P.M. Hall, Stroke surfaces: temporally coherent artistic animations from video, IEEE Trans. Vis. Comput. Graph. 11 (5) (2005) 540–549.
[27] W. Brendel, S. Todorovic, Video object segmentation by tracking regions, in: ICCV, 2009, pp. 833–840.
[28] Y. Huang, Q. Liu, D.N. Metaxas, Video object segmentation by hypergraph cut, in: CVPR, 2009, pp. 1738–1745.
[29] A.V. Reina, S. Avidan, H. Pfister, E.L. Miller, Multiple hypothesis video segmentation from superpixel flows, in: ECCV (5), 2010, pp. 268–281.
[30] M. Grundmann, V. Kwatra, M. Han, I.A. Essa, Efficient hierarchical graph-based video segmentation, in: CVPR, 2010, pp. 2141–2148.
[31] C. Xu, C. Xiong, J.J. Corso, Streaming hierarchical video segmentation, in: ECCV (6), 2012, pp. 626–639.
[32] J.Y.A. Wang, E.H. Adelson, Representing moving images with layers, IEEE Trans. Image Process. 3 (5) (1994) 625–638.
[33] D. Cremers, S. Soatto, Motion competition: A variational approach to piecewise parametric motion segmentation, Int. J. Comput. Vis. 62 (3) (2004) 249–265.
[34] A.N. Stein, T.S. Stepleton, M. Hebert, Towards unsupervised whole-object segmentation: combining automated matting with boundary detection, in: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24–26 June 2008, Anchorage, Alaska, USA, 2008.
[35] M.P. Kumar, P.H.S. Torr, A. Zisserman, Learning layered motion segmentations of video, Int. J. Comput. Vis. 76 (3) (2008) 301–319.
[36] C. Wang, M. de La Gorce, N. Paragios, Segmentation, ordering and multi-object tracking using graphical models, in: ICCV, 2009, pp. 747–754.
[37] P. Sundberg, T. Brox, M. Maire, P. Arbelaez, J. Malik, Occlusion boundary detection and figure/ground assignment from optical flow, in: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011, 2011, pp. 2233–2240.
[38] A. Ayvaci, S. Soatto, Detachable object detection: Segmentation and depth ordering from short-baseline video, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 1942–1951.
[39] A.D. Jepson, D.J. Fleet, M.J. Black, A layered motion representation with occlusion and compact spatial support, in: Computer Vision - ECCV 2002, 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28–31, 2002, Proceedings, Part I, 2002, pp. 692–706.

[40] P. Smith, T. Drummond, R. Cipolla, Layered motion segmentation and depth ordering by tracking edges, IEEE Trans. Pattern Anal. Mach. Intell. 26 (4) (2004) 479–494.
[41] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012).
[42] H. Greenspan, J. Goldberger, A. Mayer, A probabilistic framework for spatio-temporal video representation & indexing, in: ECCV, 2002, pp. 461–475.
[43] J. Wang, B. Thiesson, Y. Xu, M.F. Cohen, Image and video segmentation by anisotropic kernel mean shift, in: ECCV (2), 2004, pp. 238–249.
[44] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (6) (1984) 721–741.
[45] C. Rother, V. Kolmogorov, A. Blake, "grabcut": interactive foreground extraction using iterated graph cuts, ACM Trans. Graph. 23 (3) (2004) 309–314.
[46] Y. Boykov, G. Funka-Lea, Graph cuts and efficient N-D image segmentation, Int. J. Comput. Vis. 70 (2) (2006) 109–131.
[47] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: A survey, IEEE Trans. Med. Imaging 32 (7) (2013) 1153–1190.
[48] A. Besbes, N. Komodakis, G. Langs, N. Paragios, Shape priors and discrete MRFs for knowledge-based segmentation, in: CVPR, 2009, pp. 1295–1302.
[49] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.
[50] M.J. Wainwright, T. Jaakkola, A.S. Willsky, MAP estimation via agreement on trees: message-passing and linear programming, IEEE Trans. Inf. Theory 51 (11) (2005) 3697–3717.
[51] P. Kohli, P.H.S. Torr, Dynamic graph cuts for efficient inference in Markov random fields, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2079–2088.
[52] V. Kolmogorov, Convergent tree-reweighted message passing for energy minimization, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1568–1583.
[53] N. Komodakis, N. Paragios, G. Tziritas, MRF optimization via dual decomposition: message-passing revisited, in: IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14–20, 2007, 2007, pp. 1–8.
[54] P. Kohli, L. Ladicky, P.H.S. Torr, Robust higher order potentials for enforcing label consistency, in: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24–26 June 2008, Anchorage, Alaska, USA, 2008.
[55] N. Komodakis, N. Paragios, Beyond pairwise energies: Efficient optimization for higher-order MRFs, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20–25 June 2009, Miami, Florida, USA, 2009, pp. 2985–2992.
[56] C. Wang, O. Teboul, F. Michel, S. Essafi, N. Paragios, 3D knowledge-based segmentation using pose-invariant higher-order graphs, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010, 13th International Conference, Beijing, China, September 20–24, 2010, Proceedings, Part III, 2010, pp. 189–196.
[57] C. Wang, Y. Zeng, L. Simon, I.A. Kakadiaris, D. Samaras, N. Paragios, Viewpoint invariant 3d landmark model inference from monocular 2d images using higher-order priors, in: IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6–13, 2011, 2011, pp. 319–326.
[58] C. Wang, N. Komodakis, N. Paragios, Markov random field modeling, inference & learning in computer vision & image understanding: a survey, Comput. Vis. Image Underst. 117 (11) (2013) 1610–1627.
[59] A. Blake, P. Kohli, C. Rother, Markov Random Fields for Vision and Image Processing, MIT Press, 2011.
[60] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: ECCV, 2004, pp. 25–36.
[61] T. Jebara, J. Wang, S.-F. Chang, Graph construction and *b*-matching for semi-supervised learning, in: ICML, 2009, p. 56.
[62] E. Olson, M. Walter, J. Leonard, S. Teller, Single cluster graph partitioning for robotics applications, in: Proceedings of Robotics Science and Systems, 2005, pp. 265–272.
[63] P. Dollar, C.L. Zitnick, Structured forests for fast edge detection, in: ICCV, 2013, pp. 1841–1848.
[64] P. Arbelaez, M. Maire, C.C. Fowlkes, J. Malik, From contours to regions: an empirical evaluation, in: CVPR, 2009, pp. 2294–2301.
[65] VOT2013, The vot2013 challenge dataset, 2013, http://www.votchallenge.net (accessed 12.04.14).
[66] L. Ballan, G.J. Brostow, J. Puwein, M. Pollefeys, Unstructured video-based rendering: interactive exploration of casually captured videos, ACM Trans. Graph. 29 (4) (2010) 87:1–87:11.
[67] A. Prest, C. Leistner, J. Civera, C. Schmid, V. Ferrari, Learning object class detectors from weakly annotated video, in: CVPR, 2012, pp. 3282–3289.
[68] P. Ochs, T. Brox, Higher order motion models and spectral clustering, in: CVPR, 2012, pp. 614–621.
[69] K.D. Tang, R. Sukthankar, J. Yagnik, F. Li, Discriminative segment annotation in weakly labeled video, in: CVPR, 2013, pp. 2483–2490.
[70] Y. Zhang, X. Chen, J. Li, C. Wang, C. Xia, Semantic object segmentation via detection in weakly labeled video, in: CVPR, 2015.