SUBMODULAR VIDEO OBJECT PROPOSAL SELECTION FOR SEMANTIC OBJECT SEGMENTATION

Tinghuai Wang

Nokia Labs,
Nokia Technologies, Finland

ABSTRACT

Learning a data-driven spatio-temporal semantic representation of the objects is the key to coherent and consistent labelling in video. This paper proposes to achieve semantic video object segmentation by learning a data-driven representation which captures the synergy of multiple instances from continuous frames. To prune the noisy detections, we exploit the rich information among multiple instances and select the discriminative and representative subset. This selection process is formulated as a facility location problem solved by maximising a submodular function. Our method retrieves the longer term contextual dependencies which underpins a robust semantic video object segmentation algorithm. We present extensive experiments on a challenging dataset that demonstrate the superior performance of our approach compared with the state-of-the-art methods.

Index Terms— Submodular function, semantic video object segmentation, deep learning

1. INTRODUCTION

The proliferation of user-uploaded videos which are frequently associated with semantic tags provides a vast resource for computer vision research. These semantic tags, albeit not spatially or temporally located in the video, suggest visual concepts appearing in the video. This social trend has led to an increasing interest in exploring the idea of segmenting video objects with weak supervision or labels.

Hartmann *et al.* [1] firstly formulated the problem as learning weakly supervised classifiers for a set of independent spatio-temporal segments. Tang *et al.* [2] learned discrimative model by leveraging labelled positive videos and a large collection of negative examples based on distance matrix. Liu *et al.* [3] extended the traditional binary classifition problem to multi-class and proposed nearest neighbor-based label transfer algorithm which encourages smoothness between regions that are spatio-temporally adjacent and similar in appearance. Zhang *et al.* [4] utilized pre-trained object detector to generate a set of detections and then pruned noisy detections and regions by preserving spatio-temporal constraints.

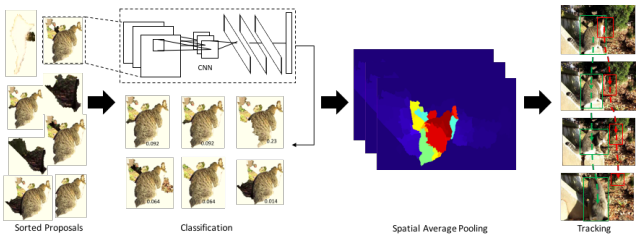


Fig. 1. An illustration of the proposed object discovery strategy.

In contrast to previous works, we propose to learn a class-specific representation which captures the synergy of multiple instances from continuous frames. To prune the noisy detections, we exploit the rich information among multiple instances and select the discriminative and representative subset. In this framework, our algorithm is able to bridge the gap between image classification and video object segmentation, leveraging the ample pre-trained image recognition models rather than strongly-trained object detectors.

2. OBJECT DISCOVERY

Semantic object segmentation requires not only localising objects of interest within a video, but also assigning class label for pixels belonging to the objects. One potential challenge of using pre-trained image recognition model to detect objects is that any regions containing the object or even part of the object, might be "correctly" recognised, which results in a large search space to accurately localise the object. To narrow down the search of targeted objects, we adopt category-independent bottom-up object proposals [5]. The proposed object discovery strategy is illustrated in Fig. 1, in which the key steps are detailed in the following sections.

2.1. Proposal Scoring and Classification

We combine the objectness score associated with each proposal from Endres and Hoiem [5] and motion information as a context cue to characterise video objects. We follow Papazoglou and Ferrari [6] which roughly produces a binary map

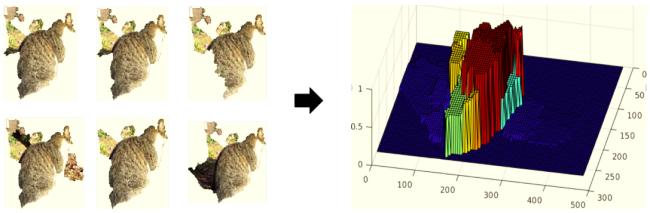


Fig. 2. An illustration of the weighted spatial average pooling strategy.

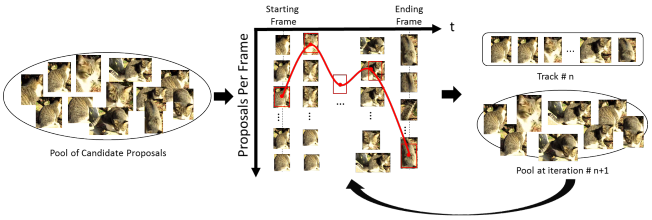


Fig. 3. Iterative tracking to eliminate spurious detections and extract consistent proposals.

indicating whether each pixel is inside the motion boundary after compensating camera motion. After acquiring the motion cues, we score each proposal r_i by both appearance and context, $s_{r_i} = \mathcal{A}(r_i) + \mathcal{C}(r_i)$, where $\mathcal{A}(r_i)$ indicates region level appearance score computed using [5] and $\mathcal{C}(r_i)$ represents the motion score of region r_i which is defined as $\mathcal{C}(r_i) = \text{Avg}(M^t(r_i)) \cdot \text{Sum}(M^t(r_i))$ where $\text{Avg}(M^t(r_i))$ and $\text{Sum}(M^t(r_i))$ compute the average and total amount of motion cues [6] included by proposal r_i on frame t respectively. Note that appearance, contextual and combined scores are normalised.

To classify each scored region proposal, we firstly warp all pixels in a bounding box around it to the required size compatible with the CNN (VGG-16 net [7] requires inputs of a fixed 224×224 pixel size), regardless its original size or shape. Prior to warping, we expand the tight bounding box by a certain number of pixels (10 in our system) around the original box, which was proven effective in the task of using image classifier for object detection task [8].

After the classification, we collect the confidence of regions with respect to the specific classes associated with the video and form a set of scored regions. For each proposal r_i, we rescore it by multiplying its score and classification confidence, which is denoted by $\tilde{s}_{r_i} = s_{r_i} \cdot c_{r_i}$.

To ensemble the multiple confidences of region proposals in each frame, we adopt a simple spatial average pooling strategy to aggregate the region-wise confidence as well as their spatial extent. For each proposal r_i, we generate confidence map $\mathcal{C}_{r_i}$ of the size of image frame, which is composited as the binary map of current region proposal multiplied by its confidence c_{r_i}. We perform an average pooling over the confidence maps of all the proposals on each frame to compute a aggregated confidence map,

$$\mathcal{C}^t = \frac{\sum_{r_i \in \mathcal{R}^t} \mathcal{C}_{r_i}}{\sum_{r_i \in \mathcal{R}^t} c_{r_i}} \tag{1}$$

where $\sum_{r_i \in \mathcal{R}^t} \mathcal{C}_{r_i}$ performs element-wise operation and $\mathcal{R}^t$ represents the set of candidate proposals from frame t.

The resulted confidence map $\mathcal{C}^t$ aggregates not only the region-wise confidence but also their spatial extent. The key insight is that good proposals coincide with each other in the spatial domain and their contribution to the final confidence map are proportional to their region-wise confidence. An illustration of the weighted spatial average pooling is shown in Fig. 2.

2.2. Tracking for Proposal Mining

Based on the computed confidence map in Eq. 1, we generate a new set of region proposals in a process analogous to the watershed algorithm, i.e., we gradually increase the threshold in defining binary maps from confidence map $\mathcal{C}^t$. This approach effectively exploit the topology structure of the confidence map. The disconnected regions thresholded at each level form the new proposals. The confidence associated with these new region proposals $\mathcal{P}$ are computed by averaging the confidence values enclosed by each region.

Due to the 2D projections, it is not possible to learn a complete representation of the object in one frame, whereas multiple image frames encompassing the same object or part of the object provide more comprehensive information. Video data naturally encodes the rich information of the objects of interest.

We propose an iterative tracking and eliminating approach to achieve these goals, as illustrated in Fig. 3. Proposals from all frames form a pool of candidates. Each iteration starts by randomly selecting a proposal on the earliest frame in the pool of candidate proposals, and it is tracked [9] until the last frame of the sequence. Any proposals whose bounding boxes with a substantial intersection-over-union (IoU) overlap (0.5 in the system) with the tracked bounding box are chosen to form a track and removed from the pool. This process iterates until the candidate pool is empty, and forms a set of tracks $\mathcal{T}$ with single-frame tracks discarded.

3. SUBMODULAR TRACK SELECTION

Given cohorts of noisy region proposals detected from video, we aim to select discriminative tracks for underpinning object segmentation. These detected tracks of region proposals may comprise false positive detections. These false positive detections may corrupt the representation learning object segmentation. Yet, pruning these sporadic false positive detections is not straightforward due to the limited recognition power of image classifier. Nonetheless, exploiting the similarities

among tracks within the same category enables recovering the discriminative subset of tracks.

For all the tracks $\mathcal{T}$ from each category, we construct a graph $\mathcal{G}_T = (\mathcal{V}, \mathcal{E})$, where each node is a track $T_i \in \mathcal{T}$ and the edges model the pairwise relations. We aim to discover a subset of tracks $\mathcal{D}$ of $\mathcal{T}$ by iteratively selecting elements of $\mathcal{T}$ into $\mathcal{D}$.

In order to obtain the discriminative tracks, we model the selection process as a facility location problem [10, 11], which can be formulated by a facility location term to compute the shared similarities and a discriminative term to preserve the discriminativity power of the selected tracks.

The facility location term is defined as,

$$\mathcal{H}(\mathcal{D}) = \sum_{i \in \mathcal{V}} \max_{j \in \mathcal{D}} w_{ij} - \sum_{i \in \mathcal{D}} \phi_i \tag{2}$$

where w_{ij} is the pairwise relation between a client node v_i and a potential facility node v_j, and ϕ_i is the cost to open a facility fixed to a constant δ. We define w_{ij} as the similarity between tracks T_i and T_j to encourage v_i to well represent or be similar to its clients, so that the final selected set $\mathcal{D}$ is representative. The weight w_{ij} is computed as:

$$w_{ij} =< F_i, F_j >, \tag{3}$$

where F_i denotes the feature vector of track T_i, which is computed by averaging L2-normalised *fc6* feature vectors from its constituent object proposals.

To enforce a class-purity constraint on the selected tracks while preserving the submodularity, a discriminative term is defined as:

$$\mathcal{P}(\mathcal{D}) = \lambda \sum_{i \in \mathcal{D}} \Phi(i), \tag{4}$$

where λ is a weight to balance the two terms, and $\Phi(i)$ denotes the averaged confidence of all constituent object proposals of track T_i.

We combine the facility location term and the discriminative term into an objective function:

$$\max_{\mathcal{D}} \mathcal{E}(\mathcal{D}) = \max_{\mathcal{D}} \mathcal{H}(\mathcal{D}) + \mathcal{P}(\mathcal{D}), \quad s.t.\ \mathcal{D} \subseteq \mathcal{T} \subseteq \mathcal{V}, N_{\mathcal{D}} \leq K, \tag{5}$$

where $N_{\mathcal{D}}$ denotes the number of open facilities and K is the number of nodes. We optimise (5) using a greedy algorithm similar to [12]. We then perform an average pooling over the score maps of all the selected tracks of proposals to compute confidence maps with respect to each category.

4. OBJECT SEGMENTATION

We formulate video object segmentation as a superpixel-labelling problem of assigning each superpixel with a label which represents background or the object class respectively. We define a space-time graph $\mathcal{G}_s = (\mathcal{V}, \mathcal{E})$ by connecting frames temporally with optical flow displacement. To achieve the efficiency and local smoothness of inference, each of the nodes in this graph is a superpixel as opposed to a pixel, and edges are set to be the spatial neighbours within the same frame and the temporal neighbours in adjacent frames connected by least one motion vector.

We define the energy function that minimises to achieve the optimal labelling:

$$E(x) = \sum_{i \in \mathcal{V}} (\psi_i^c(x_i) + \lambda_o \psi_i^o(x_i)) + \lambda_p \sum_{i \in \mathcal{V}, j \in N_i} \psi_{i,j}(x_i, x_j) \tag{6}$$

where N_i is the set of superpixels adjacent to superpixel s_i spatially and temporally in the graph respectively; λ_o and λ_p are parameters; $\psi_i^c(x_i)$ indicates the colour based unary potential and $\psi_i^o(x_i)$ is the unary potential of semantic object confidence which measures how likely the superpixel to be labelled by x_i given the semantic confidence map; $\psi_{i,j}(x_i, x_j)$ is the pairwise potential.

Colour unary potential is defined similar to [13], $\psi_i^c(x_i) = -\log U_i^c(x_i)$, where $U_i^c(\cdot)$ is the color likelihood from colour model. Two GMM colour models are estimated over the RGB values of superpixels, for objects and background respectively, by sampling the superpixel colors according to the semantic confidence map.

Semantic unary potential is defined to evaluate how likely the superpixel to be labelled by x_i given the semantic confidence map c_i^t as $\psi_i^o(x_i) = -\log U_i^o(x_i)$, where $U_i^o(\cdot)$ is the semantic likelihood, i.e., for a foreground labelling $U_i^o = c_i^t$ and $1 - c_i^t$ otherwise.

We define the pairwise potentials to encourage both spatial and temporal smoothness of labelling while preserving discontinuity in the data,

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

where $[\cdot]$ denotes the indicator function. The function $d^c(s_i, s_j)$ computes the colour distance between spatially neighbouring superpixels s_i and s_j as $d^c(s_i, s_j) = \frac{||c_i - c_j||^2}{2 < ||c_i - c_j||^2 >}$, where $||c_i - c_j||^2$ is the squared Euclidean distance between two adjacent superpixels in RGB colorspace. We adopt alpha expansion [14] to minimize Eq. 6 and the resulting label assignment gives the semantic object segmentation of the video.

5. EXPERIMENTS

In order to evaluate the performance of semantic video object segmentation, many motion segmentation or figure-ground segmentation datasets such as SegTrack [15], FBMS [16] and DAVIS [17] are not suitable due to either the ambiguous object annotation in ground-truth (one label for all foreground moving objects or no annotation for static objects) or the insufficient number of videos/frames per object class. Therefore, we evaluate our method on a large scale video dataset

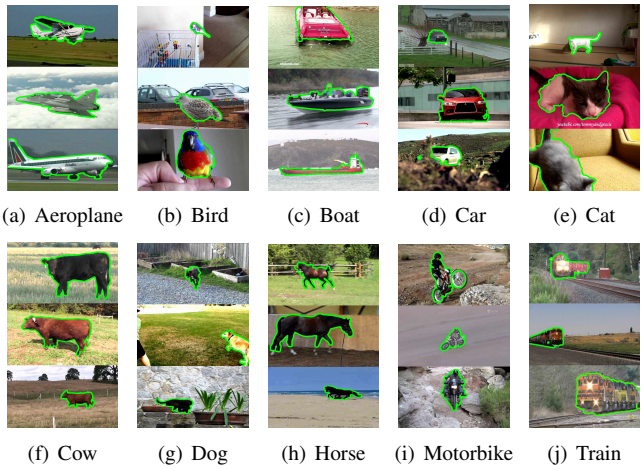
Table 1. Intersection-over-union overlap accuracies on YouTube-Objects Dataset

	LTT [20]	KOS [21]	LDW [18]	FOS [6]	DSW [2]	SSW [4]	SPS
Plane	0.539	NA	0.517	0.674	0.178	**0.758**	0.703
Bird	0.196	NA	0.175	0.625	0.198	0.608	**0.631**
Boat	0.382	NA	0.344	0.378	0.225	0.437	**0.659**
Car	0.378	NA	0.347	0.670	0.383	**0.711**	0.625
Cat	0.322	NA	0.223	0.435	0.236	0.465	**0.497**
Cow	0.218	NA	0.179	0.327	0.268	0.546	**0.701**
Dog	0.270	NA	0.135	0.489	0.237	**0.555**	0.532
Horse	0.347	NA	0.267	0.313	0.140	**0.549**	0.524
Mbike	0.454	NA	0.412	0.331	0.125	0.424	**0.554**
Train	0.375	NA	0.250	**0.434**	0.404	0.358	0.411
Cls. Avg.	0.348	0.28	0.285	0.468	0.239	0.541	**0.584**

YouTube-Objects [18]. YouTube-Objects consists of videos from 10 object classes with pixel-level ground truth for every 10 frames of 126 videos (totally more than $20,000$ frames) provided by [19] which is suitable for evaluating semantic object segmentation. These videos are very challenging and completely unconstrained, with objects of similar colour to the background, fast motion, non-rigid deformations, and fast camera motion.

We measure the segmentation performance using the standard IoU overlap as accuracy metric. We compare our approach SPS with 6 state-of-the-art automatic approaches on this dataset, including two motion driven segmentation approaches (LTT [20] and FOS [6]), three weakly supervised semantic segmentation approaches (LDW [18], DSW [2] and SSW [4]), and one object-proposal based approach (KOS [21]).

As shown in Table 1, our method surpasses the competing methods in 5 out of 10 classes, with gains up to 4.3% in category average accuracy over the best competing method SSW [4]. This is remarkable considering that SSW employed strongly-supervised deformable part models (DPM) as object detector while our approach only leverages image recognition model which lacks the capability of localising objects. SSW outperforms our method on *Car*, *Dog* and *Horse*, otherwise exhibiting varying performance across the categories — higher accuracy on fast moving objects but lower accuracy on more flexible objects *Bird*, *Cat* and slowly moving object *Cow* and *Train*. We owe it to that, though based on object detection, SSW prunes noisy detections and regions by enforcing spatio-temporal constraints, rather than learning an adapted data-driven representation in our approach. It is also worth highlighting the improvement in classes, e.g., *Cow*, where the existing methods normally fail or underperform due to the heavy reliance on motion information. The main challenge of the *Cow* videos is that cows very frequently stand still or move with mild motion, which the existing approaches might fail to capture whereas our proposed method excels by learning an object-specific representation in low-rank constrained deep feature space to retrieve the long term dependencies in the spatial-temporal domain. Our method doubles or triples



(a) Aeroplane (b) Bird (c) Boat (d) Car (e) Cat

(f) Cow (g) Dog (h) Horse (i) Motorbike (j) Train

Fig. 4. Representative successful results by our approach on YouTube-Objects dataset.

the accuracy of another weakly supervised method DSW [2] except a weaker strength on *Train* category. This is probably owing to that DSW uses a large number of similar training videos which may capture objects in rare view. Motion driven method FOS [6] can better distinguish rigid moving foreground objects on videos exhibiting relatively clean backgrounds, such as *Plane*, *Car* and *Train*. Note that, FOS is a Figure-Ground segmentation method which segments all "foreground" moving objects without assigning semantic labels.

As ablation study, we compare the full system against two baseline systems to identify the contributions of various modules: (1) system uses confidence map from scored and classified region proposals and (2) system uses confidence map from object discovery. The first baseline scheme achieves 0.527 in category/video average accuracy, whilst the second scheme improves the above accuracies by 3.4% respectively by exploiting multiple instances via an iterative tracking and eliminating strategy. By comparing the full system with the second baseline, we observe that the proposal selection algorithm contributes 2.3% respectively.

6. CONCLUSION

We have proposed a novel semantic object segmentation algorithm in weakly labeled video by harnessing the pre-trained image recognition model. Our core contribution is a video object proposal selection algorithm formulated as a submodular function, supported by the multiple instance learning via an iterative tracking and eliminating approach, which underpins a robust semantic segmentation method for unconstrained natural videos.

7. REFERENCES

[1] Glenn Hartmann, Matthias Grundmann, Judy Hoffman, David Tsai, Vivek Kwatra, Omid Madani, Sudheendra Vijayanarasimhan, Irfan A. Essa, James M. Rehg, and Rahul Sukthankar, "Weakly supervised learning of object segmentations from web-scale video," in *ECCV Workshop*, 2012, pp. 198–208.

[2] Kevin D. Tang, Rahul Sukthankar, Jay Yagnik, and Fei-Fei Li, "Discriminative segment annotation in weakly labeled video," in *CVPR*, 2013, pp. 2483–2490.

[3] Xiao Liu, Dacheng Tao, Mingli Song, Ying Ruan, Chun Chen, and Jiajun Bu, "Weakly supervised multiclass video segmentation," in *CVPR*, 2014, pp. 57–64.

[4] Yu Zhang, Xiaowu Chen, Jia Li, Chen Wang, and Changqun Xia, "Semantic object segmentation via detection in weakly labeled video," in *CVPR*, 2015, pp. 3641–3649.

[5] Ian Endres and Derek Hoiem, "Category independent object proposals," in *ECCV*, 2010, pp. 575–588.

[6] Anestis Papazoglou and Vittorio Ferrari, "Fast object segmentation in unconstrained video," in *ICCV*, 2013, pp. 1777–1784.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[9] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *ICCV*, 2015, pp. 4310–4318.

[10] Nevena Lazic, Inmar Givoni, Brendan Frey, and Parham Aarabi, "Floss: Facility location for subspace segmentation," in *CVPR*. IEEE, 2009, pp. 825–832.

[11] Yi-Hsuan Tsai, Guangyu Zhong, and Ming-Hsuan Yang, "Semantic co-segmentation in videos," in *ECCV*, 2016, pp. 760–775.

[12] Fan Zhu, Zhuolin Jiang, and Ling Shao, "Submodular object recognition," in *CVPR*, 2014, pp. 2457–2464.

[13] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[14] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

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

[16] Peter Ochs, Jitendra Malik, and Thomas Brox, "Segmentation of moving objects by long term video analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1187–1200, 2014.

[17] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016.

[18] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari, "Learning object class detectors from weakly annotated video," in *CVPR*, 2012, pp. 3282–3289.

[19] Suyog Dutt Jain and Kristen Grauman, "Supervoxel-consistent foreground propagation in video," in *ECCV*, pp. 656–671. Springer, 2014.

[20] Thomas Brox and Jitendra Malik, "Object segmentation by long term analysis of point trajectories," in *ECCV*, 2010, pp. 282–295.

[21] Yong Jae Lee, Jaechul Kim, and Kristen Grauman, "Key-segments for video object segmentation," in *ICCV*, 2011, pp. 1995–2002.