Semi-supervised Domain Adaptation for Weakly Labeled Semantic Video Object Segmentation

Huiling Wang[1](✉), Tapani Raiko[1], Lasse Lensu[2], Tinghuai Wang[3], and Juha Karhunen[1]

[1] Aalto University, Espoo, Finland
huling.wang@tut.fi
[2] Lappeenranta University of Technology, Lappeenranta, Finland
[3] Nokia Technologies, Tampere, Finland

Abstract. Deep convolutional neural networks (CNNs) have been immensely successful in many high-level computer vision tasks given large labelled datasets. However, for video semantic object segmentation, a domain where labels are scarce, effectively exploiting the representation power of CNN with limited training data remains a challenge. Simply borrowing the existing pre-trained CNN image recognition model for video segmentation task can severely hurt performance. We propose a semi-supervised approach to adapting CNN image recognition model trained from labelled image data to the target domain exploiting both semantic evidence learned from CNN, and the intrinsic structures of video data. By explicitly modelling and compensating for the domain shift from the source domain to the target domain, this proposed approach underpins a robust semantic 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

Semantically assigning each pixel in video with a known class label can be challenging for machines due to several reasons. Firstly, acquiring the prior knowledge about object appearance, shape or position is difficult. Secondly, gaining pixel-level annotation for training supervised learning algorithms is prohibitively expensive comparing with image-level labelling. Thirdly, background clutters, occlusion and object appearance variations introduce visual ambiguities that in turn induce instability in boundaries and the potential for localised under- or over-segmentation. Recent years have seen encouraging progress, particularly in terms of generic object segmentation [1–6], and the success of convolutional neural networks (CNNs) in image recognition [7–9] also sheds light on semantic video object segmentation.

Generic object segmentation methods [2,3,5,10–12] largely utilise category independent region proposal methods [13,14], to capture object-level description

S.-H. Lai et al. (Eds.): ACCV 2016, Part I, LNCS 10111, pp. 163–179, 2017.
DOI: 10.1007/978-3-319-54181-5_11

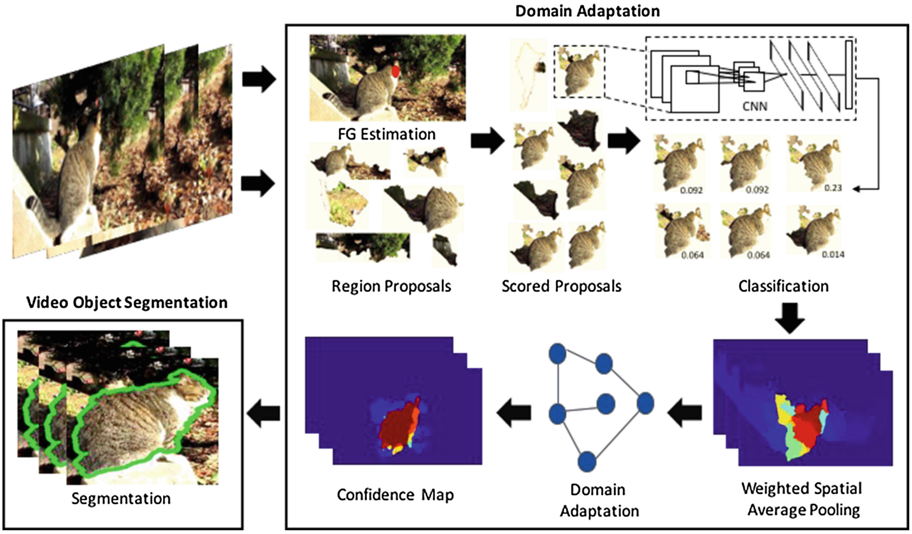


Fig. 1. Overview of our proposed method.

of the generic object in the scene incorporating motion cues. These approaches address the challenge of visual ambiguities to some extent, seeking the weak prior knowledge of what the object may look like and where it might be located. However, there are generally two major issues with these approaches. Firstly, the generic detection has very limited capability to determine the presence of an object. Secondly, such approaches are generally unable to determine and differentiate unique multiple objects, regardless of categories. These two bottlenecks limit these approaches to segmenting one single object or all foreground objects regardless classes or identifies.

Deep convolutional neural networks have been proven successful [7–9] in many high-level computer vision tasks such as image recognition and object detection. However, stretching this success to the domain of pixel-level classification or labelling, i.e., semantic segmentation, is not naturally straightforward. This is not only owing to the difficulties of collecting pixel-level annotations, but also due to the nature of large receptive fields of convolutional neural networks. Furthermore, the aforementioned challenges present in video data demand a data-driven representation of the video object in order to give a spatio-temporal coherent segmentation. This motivates us to develop a framework for adapting image recognition models (e.g., CNN) trained on static images to a video domain for the demanding task of pixel labelling. This goal is achieved by proposing a semi-supervised domain adaptation approach to forming a data-driven object representation which incorporates both the semantic evidence from pre-trained CNN image recognition model and the constraint imposed by the intrinsic structure of video data. We exploit the constraint in video data that when the same object is recurring between video frames, the spatio-temporal coherence implies the associated unlabelled data to be the same label. This data-driven object representation underpins a robust object segmentation method for weakly labelled natural videos.

The paper is structured as follows: We firstly review related work in video object segmentation (Sect. 2). Our method introduced in Sects. 3 and 4 consists of domain adaptation and segmentation respectively, as shown in Fig. 1. Evaluations and comparisons in Sect. 5 show the benefits of our method. We conclude this paper with our findings in Sect. 6.

2 Related Work

Video object segmentation has received considerable attention in recent years, with the majority of research effort categorised into three groups based on the level of supervisions: (semi-)supervised, unsupervised and weakly supervised methods.

Methods in the first category normally require an initial annotation of the first frame, which either perform spatio-temporal grouping [15,16] or propagate the annotation to drive the segmentation in successive frames [17–20].

Unsupervised methods have been proposed as a consequence of the prohibitive cost of human-in-the-loop operations when processing ever-growing large-scale video data. Bottom-up approaches [4,21,22] largely utilise spatio-temporal appearance and motion constraints, while motion segmentation approaches [23,24] perform long-term motion analysis to cluster pixels or regions in video data. Giordano *et al.* [25] extended [4] by introducing 'perceptual organisation' to improve segmentation. Taylor *et al.* [26] inferred object segmentation through long-term occlusion relations, and introduced a numerical scheme to perform partition directly on pixel grid. Wang *et al.* [27] exploited saliency measure using geodesic distance to build global appearance models. Several methods [2,3,5,6,11] propose to introduce a top-down notion of object by exploring recurring object-like regions from still images by measuring generic object appearance (e.g., [13]) to achieve state-of-the-art results. However, due to the limited recognition capability of generic object detection, these methods normally can only segment foreground objects regardless of semantic label.

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.* [28] firstly formulated the problem as learning weakly supervised classifiers for a set of independent spatio-temporal segments. Tang *et al.* [29] learned discriminative model by leveraging labelled positive videos and a large collection of negative examples based on distance matrix. Liu *et al.* [30] extended the traditional binary classification problem to multi-class and proposed nearest-neighbour-based label transfer algorithm which encourages smoothness between regions that are spatio-temporally adjacent and similar in appearance. Zhang *et al.* [31] utilised pre-trained object detector to generate a set of detections and then pruned noisy detections and regions by preserving spatio-temporal constraints.

3 Domain Adaptation

We set out our approach to first semantically discovering possible objects of interest from video. We then adapt the source domain from image recognition to the target domain, i.e., pixel or superpixel level labelling. This approach is built by additionally incorporating constraints obtained from a given similarity graph defined on unlabelled target instances.

3.1 Object Discovery

Proposal Scoring. Unlike image classification or object detection, semantic object segmentation requires not only localising objects of interest within an image, but also assigning class label for pixels belonging to the objects. One potential challenge of using image classifier 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.

As we are interested in producing segmentations and not just bounding boxes, we require region proposals. We consider those regions as candidate object hypotheses. The objectness score associated with each proposal from [13] indicates how likely it is for an image region contain an object of any class. However, 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.). We incorporate motion information as a context cue for video objects. There has been many previous works on estimating local motion cues and we adopt a motion boundary based approach as introduced in [4] which roughly produces a binary map indicating whether each pixel is inside the motion boundary after compensating camera motion. After acquiring the motion cues, we score each proposal r by both appearance and context,

$$s_r = \mathcal{A}(r) + \mathcal{C}(r)$$

where $\mathcal{A}(r)$ indicates region level appearance score computed using [13] and $\mathcal{C}(r)$ represents the contextual score of region r which is defined as:

$$\mathcal{C}(r) = \mathrm{Avg}(M^t(r)) \cdot \mathrm{Sum}(M^t(r))$$

where $\mathrm{Avg}(M^t(r))$ and $\mathrm{Sum}(M^t(r))$ compute the average and total amount of motion cues [4] included by proposal r on frame t respectively. Note that appearance, contextual and combined scores are normalised.

Proposal Classification. On each frame t we have a collection of region proposals scored by their appearance and contextual information. These region proposals may contain various objects present in the video. In order to identify the objects of interest specified by the video level tag, region level classification is performed. We consider proven classification architectures such as VGG-16 nets [8] which did exceptionally well in ILSVRC14. VGG-16 net uses 3×3 convolution interleaved with max pooling and 3 fully-connected layers.

In order to classify each region proposal, we firstly warp the image data in each region into a form that is compatible with the CNN (VGG-16 net requires inputs of a fixed 224×224 pixel size). Although there are many possible transformations of our arbitrary-shaped regions, we warp all pixels in a bounding box around it to the required 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 [32].

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,

$$\{\mathcal{H}_{w_1}, \ldots, \mathcal{H}_{w_K}\}$$

where

$$\mathcal{H}_{w_k} = \{(r_1, s_{r_1}, c_{r_1,w_k}), \ldots, (r_N, s_{r_N}, c_{r_N,w_k})\}$$

with s_{r_i} is the original score of proposal r_i and c_{r_i,w_k} is its confidence from CNN classification with regard to keyword or class w_k. Figure 1 shows the positive detections with confidence higher than a predefined threshold (0.01), where higher confidence does not necessarily correspond to good proposals. This is mainly due to the nature of image classification where the image frame is quite often much larger than the tight bounding box of the object. In the following discussion we drop the subscript of classes, and formulate our method with regard to one single class for the sake of clarity, albeit our method works on multiple classes.

Spatial Average Pooling. After the initial discovery, a large number of region proposals are positively detected with regard to a class label, which include overlapping regions on the same objects and spurious detections. We adopt a simple weighted spatial average pooling strategy to aggregate the region-wise score, confidence as well as their spatial extent. For each proposal r_i, we rescore it by multiplying its score and classification confidence, which is denoted by $\tilde{s}_{r_i} = s_{r_i} \cdot c_{r_i}$. We then generate score map $\mathcal{S}_{r_i}$ of the size of image frame, which is composited as the binary map of current region proposal multiplied by its score $\tilde{s}_{r_i}$. We perform an average pooling over the score maps of all the proposals to compute a confidence map,

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

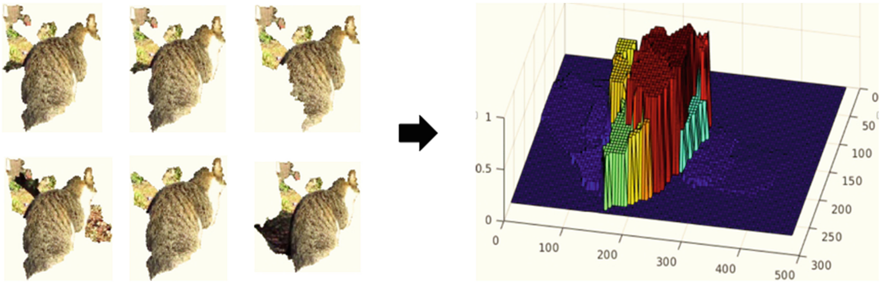


Fig. 2. An illustration of the weighted spatial average pooling strategy.

where $\sum_{r_i \in \mathcal{R}^t} \mathcal{S}_{r_i}$ 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 score 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 score. An illustration of the weighted spatial average pooling is shown in Fig. 2.

3.2 Semi-supervised Domain Adaptation

To perform domain adaptation from image recognition to video object segmentation, we define a weighted space-time graph $\mathcal{G}_d = (\mathcal{V}_d, \mathcal{E}_d)$ spanning the whole video or a shot with each node corresponding to a superpixel, and each edge connecting two superpixels based on spatial and temporal adjacencies. Temporal adjacency is coarsely determined based on motion estimates, i.e., two superpixels are deemed temporally adjacent if they are connected by at least one motion vector.

We compute the affinity matrix A of the graph among spatial neighbours as

$$A^s_{i,j} = \frac{\exp(-d^c(s_i, s_j))}{d^s(s_i, s_j)} \tag{2}$$

where the functions $d^s(s_i, s_j)$ and $d^c(s_i, s_j)$ computes the spatial and color distances between spatially neighbouring superpixels s_i and s_j respectively:

$$d^c(s_i, s_j) = \frac{||c_i - c_j||^2}{2 < ||c_i - c_j||^2 >}$$

where $||c_i - c_j||^2$ is the squared Euclidean distance between two adjacent superpixels in RGB colour space, and $< \cdot >$ computes the average over all pairs i and j.

For affinities among temporal neighbours s_i^{t-1} and s_j^t, we consider both the temporal and colour distances between s_i^{t-1} and s_j^t,

$$A^t_{i,j} = \frac{\exp(-d^c(s_i, s_j))}{d^t(s_i, s_j)}$$

where

$$\begin{aligned} d^t(s_i, s_j) &= \frac{1}{m_i \cdot \rho_{i,j}}, \\ m_i &= \exp(-w_c \cdot \pi_i), \\ \rho_{i,j} &= \frac{|\tilde{s}_i^{t-1} \cap s_j^t|}{|\tilde{s}_i^{t-1}|}. \end{aligned} \tag{3}$$

Specifically, we define the temporal distance $d^t(s_i, s_j)$ by combining two factors, i.e., the temporal overlapping ratio $\rho_{i,j}$ and motion accuracy m_i. π_i denotes the motion coherence, and $w_c = 2.0$ is a parameter. The larger the temporal overlapping ratio is between two temporally related superpixels, the closer they are in temporal domain, subject to the accuracy of motion estimation. The temporal overlapping ratio $\rho_{i,j}$ is defined between the warped version of s_i^{t-1} following motion vectors and s_j^t, where $\tilde{s}_i^{t-1}$ is the warped region of s_i^{t-1} by optical flow to frame t, and $|\cdot|$ is the cardinality of a superpixel. The reliability of motion estimation inside s_i^{t-1} is measured by the motion coherence. A superpixel, i.e., a small portion of a moving object, normally exhibits coherent motions. We correlate the reliability of motion estimation of a superpixel with its local motion coherence. We compute quantised optical flow histograms h_i for superpixel s_i^{t-1}, and compute π_i as the information entropy of h_i. Smaller π_i indicates higher levels of motion coherence, i.e., higher motion reliability of motion estimation. An example of computed motion reliability map is shown in Fig. 3.

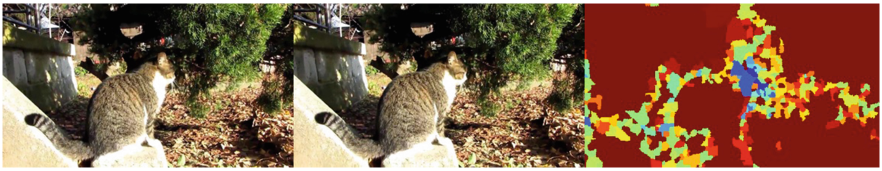


Fig. 3. Motion reliability map (right) computed given the optical flow between two consecutive frames (left and middle).

We follow a similar formulation with [33] to minimise an energy function $E(X)$ with respect to all superpixels confidence X $(X \in [-1, 1])$:

$$E(X) = \sum_{i,j=1}^{N} A_{ij} ||x_i d_i^{-\frac{1}{2}} - x_j d_j^{-\frac{1}{2}}||^2 + \mu \sum_{i=1}^{N} ||x_i - c_i||^2, \tag{4}$$

where μ is the regularisation parameter, and X are the desirable confidence of superpixels which are imposed by noisy confidence C in Eq. (1). We set $\mu = 0.5$. Let the node degree matrix $D = \text{diag}([d_1, \ldots, d_N])$ be defined as $d_i = \sum_{j=1}^{N} A_{ij}$, where $N = |\mathcal{V}|$. Denoting $S = D^{-1/2} A D^{-1/2}$, this energy function can be minimised iteratively as

$$X^{t+1} = \alpha S X^t + (1-\alpha)C$$

until convergence, where α controls the relative amount of the confidence from its neighbours and its initial confidence. Specifically, the affinity matrix A of $\mathcal{G}_d$ is symmetrically normalised in S, which is necessary for the convergence of the following iteration. In each iteration, each superpixel adapts itself by receiving the confidence from its neighbours while preserving its initial confidence. The confidence is adapted symmetrically since S is symmetric. After convergence, the confidence of each unlabelled superpixel is adapted to be the class of which it has received most confidence during the iterations (Fig. 4).

We alternatively solve the optimisation problem as a linear system of equations which is more efficient. Differentiating $E(X)$ with respect to X we have

$$\nabla E(X)|_{X=X^*} = X^* - SX^* + \mu(X^* - C) = 0 \tag{5}$$

which can be transformed as

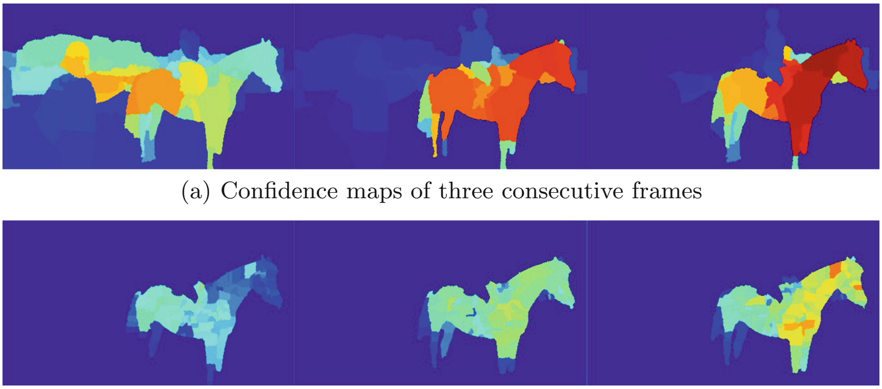
$$(I - (1 - \frac{\mu}{1+\mu})S)X^* = \frac{\mu}{1+\mu}C. \tag{6}$$

Finally we have

$$(I - (1-\eta)S)X^* = \eta C. \tag{7}$$

where $\eta = \frac{\mu}{1+\mu}$.

The optimal solution for X can be found using the preconditioned (Incomplete Cholesky factorisation) conjugate gradient method with very fast convergence. For consistency, still let C denote the optimal semantic confidence X for the rest of this paper.



(a) Confidence maps of three consecutive frames

(b) Confidence maps after domain adaptation

Fig. 4. Proposed domain adaptation effectively adapts the noisy confidence map from image recognition to the video object segmentation domain.

4 Video Object Segmentation

We formulate video object segmentation as a superpixel-labelling problem of assigning each superpixel two classes: objects and background (not listed in the keywords). Similar to Subsect. 3.2 we define a space-time superpixel graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$ by connecting frames temporally with optical flow displacement.

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

$$E(x) = \sum_{i \in \mathcal{V}} (\psi_i^c(x_i) + \lambda_o \psi_i^o(x_i)) + \lambda_s \sum_{i \in \mathcal{V}, j \in N_i^s} \psi_{i,j}^s(x_i, x_j) + \lambda_t \sum_{i \in \mathcal{V}, j \in N_i^t} \psi_{i,j}^t(x_i, x_j) \tag{8}$$

where N_i^s and N_i^t are the sets of superpixels adjacent to superpixel s_i spatially and temporally in the graph respectively; λ_o, λ_s and λ_t are parameters; $\psi_i^c(x_i)$ indicates the color 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; $\psi_{i,j}^s(x_i, x_j)$ and $\psi_{i,j}^t(x_i, x_j)$ are spatial pairwise potential and temporal pairwise potential respectively. We set parameters $\lambda_o = 10$, $\lambda_s = 1000$ and $\lambda_t = 2000$. The definitions of these unary and pairwise terms are explained in detail next.

4.1 Unary Potentials

We define unary terms to measure how likely a superpixel is to be label as background or the object of interest according to both the appearance model and semantic object confidence map.

Colour unary potential is defined similar to [34], which evaluates the fit of a colour distribution (of a label) to the colour of a superpixel,

$$\psi_i^c(x_i) = -\log U_i^c(x_i)$$

where $U_i^c(\cdot)$ is the colour likelihood from colour model.

We train two Gaussian Mixture Models (GMMs) over the RGB values of superpixels, for objects and background respectively. These GMMs are estimated by sampling the superpixel colours 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

$$\psi_i^o(x_i) = -\log U_i^o(x_i)$$

where $U_i^o(\cdot)$ is the semantic likelihood, i.e., for an object labelling $U_i^o = c_i^t$ and $1 - c_i^t$ otherwise.

4.2 Pairwise Potentials

We define the pairwise potentials to encourage both spatial and temporal smoothness of labelling while preserving discontinuity in the data. These terms are defined similar to the affinity matrix in Subsect. 3.2.

Superpixels in the same frame are spatially connected if they are adjacent. The spatial pairwise potential $\psi^s_{i,j}(x_i, x_j)$ penalises different labels assigned to spatially adjacent superpixels:

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

where $[\cdot]$ denotes the indicator function.

The temporal pairwise potential is defined over edges where superpixels are temporally connected on consecutive frames. Superpixels s_i^{t-1} and s_j^t are deemed as temporally connected if there is at least one pixel of s_i^{t-1} which is propagated to s_j^t following the optical flow motion vectors,

$$\psi^t_{i,j}(x_i, x_j) = \frac{[x_i \neq x_j]\exp(-d^c(s_i, s_j))}{d^t(s_i, s_j)}.$$

Taking advantage of the similar definitions in computing affinity matrix in Subsect. 3.2, the pairwise potentials can be efficiently computed by reusing the affinity in Eqs. (2) and (3).

4.3 Optimisation

We adopt alpha expansion [35] to minimise Eq. (8) and the resulting label assignment gives the semantic object segmentation of the video.

4.4 Implementation

We implement our method using MATLAB and C/C++, with Caffe [36] implementation of VGG-16 net [8]. We reuse the superpixels returned from [13] which is produced by [37]. Large displacement optical flow algorithm [38] is adopted to cope with strong motion in natural videos. 5 components per GMM in RGB colour space are learned to model the colour distribution following [34]. Our domain adaptation method performs efficient learning on superpixel graph with an unoptimised MATLAB/C++ implementation, which takes around 30 s over a video shot of 100 frames. The average time on segmenting one preprocessed frame is about 3 s on a commodity desktop with a Quad-Core 4.0 GHz processor, 16 GB of RAM, and GTX 980 GPU.

We set parameters by optimising segmentation against ground truth over a sampled set of 5 videos from publicly available *Freiburg-Berkeley Motion Segmentation Dataset* dataset [39] which proved to be a versatile setting for a wide variety of videos. These parameters are fixed for the evaluation.

5 Evaluation

We evaluate our method on a large scale video dataset YouTube-Objects [40] and SegTrack [18]. YouTube-Objects consists of videos from 10 object classes with pixel-level ground truth for every 10 frames of 126 videos provided by [41]. 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. SegTrack consists of 5 videos with single or interacting objects presented in each video.

5.1 YouTube-Objects Dataset

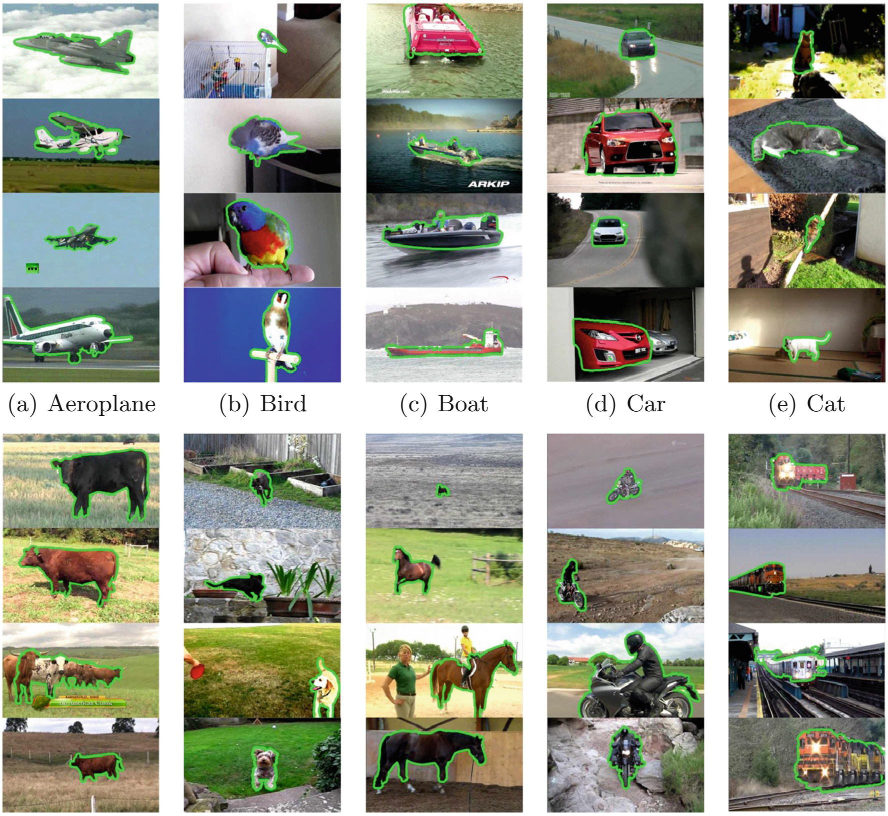
We measure the segmentation performance using the standard *intersection-over-union* (IoU) overlap as accuracy metric. We compare our approach with 6 state-of-the-art automatic approaches on this dataset, including two motion driven segmentation [1,4], three weakly supervised approaches [29,31,40], and state-of-the-art object-proposal based approach [2]. Among the compared approaches, [1,2] reported their results by fitting a bounding box to the largest connected segment and overlapping with the ground-truth bounding box; the result of [2] on this dataset is originally reported by [4] by testing on 50 videos (5/class). The performance of [4] measured with respect to segmentation ground-truth is reported by [31]. Zhang *et al.* [31] reported results in more than 5500 frames sampled in the dataset based on the segmentation ground-truth. Wang *et al.* [27] reported the average results on 12 randomly sampled videos in terms of a different metric, i.e., per-frame pixel errors across all categories, and thus not listed here for comparison.

As shown in Table 1 and Fig. 5, our method outperforms the competing methods in 7 out of 10 classes, with gains up to 6.3%/6.6% in category/video average accuracy over the best competing method [31]. This is remarkable considering that [31] employed strongly-supervised deformable part models (DPM)

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

	Brox [1]	Lee [2]	Prest [40]	Papazoglou [4]	Tang [29]	Zhang [31]	Baseline	Ours
Plane	0.539	NA	0.517	0.674	0.178	**0.758**	0.693	0.757
Bird	0.196	NA	0.175	0.625	0.198	0.608	0.590	**0.658**
Boat	0.382	NA	0.344	0.378	0.225	0.437	0.564	**0.656**
Car	0.378	NA	0.347	0.670	0.383	**0.711**	0.594	0.650
Cat	0.322	NA	0.223	0.435	0.236	0.465	0.455	**0.514**
Cow	0.218	NA	0.179	0.327	0.268	0.546	0.647	**0.714**
Dog	0.270	NA	0.135	0.489	0.237	0.555	0.495	**0.570**
Horse	0.347	NA	0.267	0.313	0.140	0.549	0.486	**0.567**
Mbike	0.454	NA	0.412	0.331	0.125	0.424	0.480	**0.560**
Train	0.375	NA	0.250	**0.434**	0.404	0.358	0.353	0.392
Cls. Avg	0.348	0.28	0.285	0.468	0.239	0.541	0.536	**0.604**
Vid. Avg	NA	NA	NA	0.432	0.228	0.526	0.523	**0.592**

as object detector while our approach only leverages image recognition model which lacks the capability of localising objects. [31] outperforms our method on *Plane* and *Car*, otherwise exhibiting varying performance across the categories — higher accuracy on more rigid objects but lower accuracy on highly flexible and deformable objects such as *Cat* and *Dog*. We owe it to that, though based on object detection, [31] 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 remarking on 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 leveraging the recognition and representation power of deep convolutional neural network, as well as the semi-supervised domain adaptation.



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

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

Fig. 5. Representative successful results by our approach on YouTube-Objects dataset.

Interestingly, another weakly supervised method [29] slightly outperforms our method on *Train* although all methods do not perform very well on this category due to the slow motion and missed detections on partial views of trains. This is probably owing to that [29] uses a large number of similar training videos which may capture objects in rare view. Otherwise, our method doubles or triples the accuracy of [29]. Motion driven method [4] can better distinguish rigid moving foreground objects on videos exhibiting relatively clean backgrounds, such as *Plane* and *Car*.

As ablation study, we evaluate a baseline scheme by removing the proposed domain adaptation algorithm (Sect. 3.2) from the full system. As shown in Table 1, the proposed semi-supervised domain adaptation is able to learn to successfully adapt to the target with a gain of 6.8%/6.9% in category/video average accuracies, comparing with the baseline scheme using only the semantic confidence by merging initially discovered region proposals (Sect. 3.1) for segmentation (with accuracies 0.536/0.523). This adaptation from the source domain of image recognition to the target domain of video semantic segmentation effectively compensates for the paradigm shift which is the key of our proposed method to outperform the state-of-the-art despite the use of weakly supervised image classifier.

5.2 SegTrack Dataset

We evaluate on SegTrack dataset to focus our comparison with the state-of-the-art semantic object segmentation algorithm [31] driven by object detector. We also compare with co-segmentation method [42] and the representative Figure-Ground segmentation algorithms [1–4, 27, 31] as baselines. To avoid confusion of segmentation results, all the compared methods only consider the primary object.

As shown in Table 2, our method outperforms the semantic segmentation [31] on *birdfall* and *monkeydog* videos, motion driven method [4] on four out of five videos, proposal ranking method [2] on four videos, proposal merging method [3] and saliency driven method [27] on two videos respectively. Clustering point tracks based method [1] results in highest error among all the methods. Co-segmentation method [42] reported the state-of-the-art results on three out of

Table 2. Quantitative segmentation results on SegTrack. Segmentation error as measured by the average number of incorrect pixels per frame.

Video (No. frames)	Ours	[1]	[4]	[3]	[2]	[42]	[31]	[27]
birdfall (30)	170	468	217	155	288	152	339	209
cheetah (29)	826	1968	890	633	905	NA	803	796
girl (21)	1647	7595	3859	1488	1785	1053	1459	1040
monkeydog (71)	304	1434	284	472	521	NA	365	562
parachute (51)	363	1113	855	220	201	189	196	207

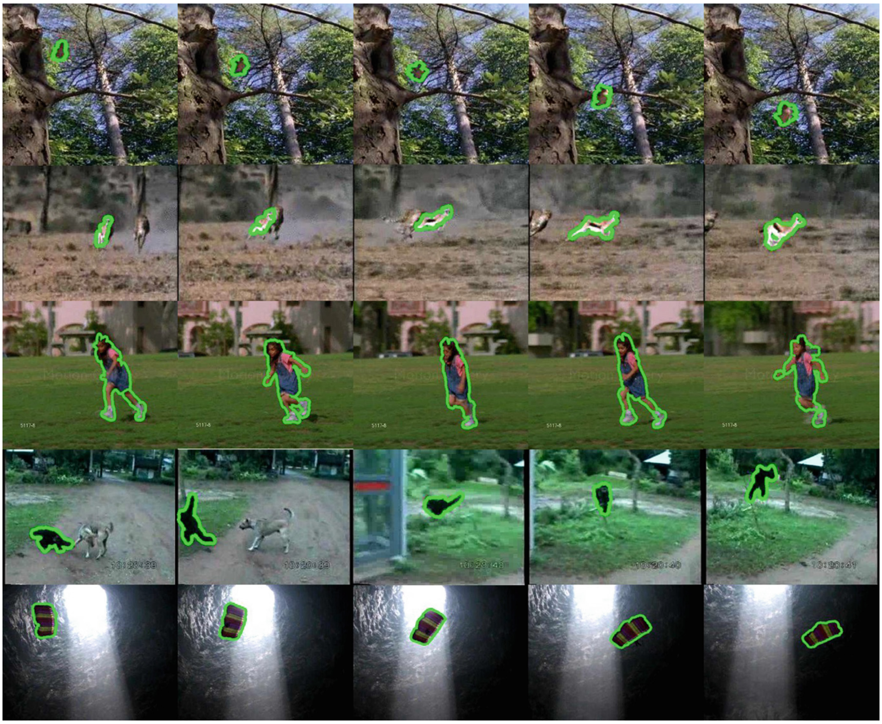


Fig. 6. Qualitative results of our method on SegTrack dataset.

five videos from SegTrack, albeit it can only segment single object as opposed to our method which can deal with objects of multiple semantic categories. Overall, our performance is about on par with the state-of-the-art semantic object segmentation method [31]. Qualitative segmentation of our approach is shown in Fig. 6.

6 Conclusion

We have proposed a semi-supervised framework to adapt CNN classifiers from image recognition domain to the target domain of semantic video object segmentation. This framework combines the recognition and representation power of CNN with the intrinsic structure of unlabelled data in the target domain to improve inference performance, imposing spatio-temporal smoothness constraints on the semantic confidence over the unlabelled video data. This proposed domain adaptation framework enables learning a data-driven representation of video objects. We demonstrated that this representation underpins a robust semantic video object segmentation method which outperforms existing methods on challenging datasets. As a future work, it would be interesting to incorporate representations learned from higher layers of CNN into the domain adaptation, which might potentially improve adaptation by propagating and combining higher level context.

References

1. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 282–295. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15555-0_21
2. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV, pp. 1995–2002 (2011)
3. Zhang, D., Javed, O., Shah, M.: Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In: CVPR, pp. 628–635 (2013)
4. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV, pp. 1777–1784 (2013)
5. Wang, T., Wang, H.: Graph transduction learning of object proposals for video object segmentation. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9006, pp. 553–568. Springer, Heidelberg (2015). doi:10.1007/978-3-319-16817-3_36
6. Wang, H., Wang, T.: Primary object discovery and segmentation in videos via graph-based transductive inference. Comput. Vis. Image Underst. **143**, 159–172 (2016)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1106–1114 (2012)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint (2014). arXiv:1409.1556
9. Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T.: Semi-supervised learning with ladder network. In: NIPS (2015)
10. Faktor, A., Irani, M.: Video segmentation by non-local consensus voting. In: BMVC, vol. 2, p. 6 (2014)
11. Yang, J., Zhao, G., Yuan, J., Shen, X., Lin, Z., Price, B., Brandt, J.: Discovering primary objects in videos by saliency fusion and iterative appearance estimation. IEEE Trans. Circuits Syst, Video Technol (2015)
12. Perazzi, F., Wang, O., Gross, M., Sorkine-Hornung, A.: Fully connected object proposals for video segmentation. In: ICCV, pp. 3227–3234 (2015)
13. Endres, I., Hoiem, D.: Category independent object proposals. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 575–588. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15555-0_42
14. Manen, S., Guillaumin, M., Gool, L.J.V.: Prime object proposals with randomized prim's algorithm. In: ICCV, pp. 2536–2543 (2013)
15. Wang, J., Xu, Y., Shum, H.Y., Cohen, M.F.: Video tooning. ACM Trans. Graph. **23**, 574–583 (2004)
16. Collomosse, J.P., Rowntree, D., Hall, P.M.: Stroke surfaces: temporally coherent artistic animations from video. IEEE Trans. Vis. Comput. Graph. **11**, 540–549 (2005)
17. Wang, T., Collomosse, J.P.: Probabilistic motion diffusion of labeling priors for coherent video segmentation. IEEE Trans. Multimed. **14**, 389–400 (2012)

18. Tsai, D., Flagg, M., Nakazawa, A., Rehg, J.M.: Motion coherent tracking using multi-label MRF optimization. Int. J. Comput. Vis. **100**, 190–202 (2012)
19. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV, Australia, 1–8 December 2013, pp. 2192–2199 (2013)
20. Wang, T., Han, B., Collomosse, J.P.: Touchcut: fast image and video segmentation using single-touch interaction. Comput. Vis. Image Underst. **120**, 14–30 (2014)
21. Grundmann, M., Kwatra, V., Han, M., Essa, I.A.: Efficient hierarchical graph-based video segmentation. In: CVPR, pp. 2141–2148 (2010)
22. Xu, C., Xiong, C., Corso, J.J.: Streaming hierarchical video segmentation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 626–639. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_45
23. Wang, C., de La Gorce, M., Paragios, N.: Segmentation, ordering and multi-object tracking using graphical models. In: ICCV, pp. 747–754 (2009)
24. Sundberg, P., Brox, T., Maire, M., Arbelaez, P., Malik, J.: Occlusion boundary detection and figure/ground assignment from optical flow. In: CVPR, pp. 2233–2240 (2011)
25. Giordano, D., Murabito, F., Palazzo, S., Spampinato, C.: Superpixel-based video object segmentation using perceptual organization and location prior. In: CVPR, pp. 4814–4822 (2015)
26. Taylor, B., Karasev, V., Soatto, S.: Causal video object segmentation from persistence of occlusions. In: CVPR, pp. 4268–4276 (2015)
27. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: CVPR, pp. 3395–3402 (2015)
28. Hartmann, G., Grundmann, M., Hoffman, J., Tsai, D., Kwatra, V., Madani, O., Vijayanarasimhan, S., Essa, I., Rehg, J., Sukthankar, R.: Weakly supervised learning of object segmentations from web-scale video. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 198–208. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_20
29. Tang, K.D., Sukthankar, R., Yagnik, J., Li, F.: Discriminative segment annotation in weakly labeled video. In: CVPR, pp. 2483–2490 (2013)
30. Liu, X., Tao, D., Song, M., Ruan, Y., Chen, C., Bu, J.: Weakly supervised multi-class video segmentation. In: CVPR, pp. 57–64 (2014)
31. Zhang, Y., Chen, X., Li, J., Wang, C., Xia, C.: Semantic object segmentation via detection in weakly labeled video. In: CVPR, pp. 3641–3649 (2015)
32. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587 (2014)
33. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Sch, B.: Learning with local and global consistency. In: NIPS, pp. 321–328 (2004)
34. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23**, 309–314 (2004)
35. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 1222–1239 (2001)
36. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
37. Arbelaez, P., Maire, M., Fowlkes, C.C., Malik, J.: From contours to regions: an empirical evaluation. In: CVPR, pp. 2294–2301 (2009)

38. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24673-2_3
39. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 282–295. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15555-0_21
40. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR, pp. 3282–3289 (2012)
41. Jain, S.D., Grauman, K.: Supervoxel-consistent foreground propagation in video. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 656–671. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_43
42. Wang, L., Hua, G., Sukthankar, R., Xue, J., Zheng, N.: Video object discovery and co-segmentation with extremely weak supervision. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 640–655. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_42