Multiclass Semantic Video Segmentation with Object-level Active Inference

Buyu Liu, Xuming He
Australian National University, National ICT of Australia

Semantic scene parsing has recently made much progress by incorporating high-level visual information, such as scene context and objects, and jointly solving multiple related vision tasks [5, 7]. However, such object-aware strategies require many proposals of object instances and their relations to accommodate uncertainty in object detection and localization. This leads to increasingly higher complexity of the resulting models on pixels and objects, which is challenging for efficient inference at test time, especially for parsing video data.

In this work, we address the problem of integrating object reasoning with supervoxel labeling in multiclass semantic video segmentation. Taking a hypothesize-and-verify approach, we generate a pool of object proposals and formulate the video segmentation as a joint labeling of pixels and object hypotheses. To handle a large number of object proposals, we adopt an active inference strategy at object level to select an optimal subset of proposals for joint inference. An overview of our model is shown in Figure 1.

We tackle the joint labeling problem by designing an object-augmented dense CRF in spatio-temporal domain, which captures long-range dependency between supervoxels, and imposes consistency between object and supervoxel labels. Specifically, given a video sequence $\mathcal{T}$, we first compute its supervoxel representation and the semantic class of the ith supervoxel is denoted as l_i. We then generate a set of object trajectory proposals from object detection and tracking efficiently as in [4]. For the m-th proposal, a binary variable d_m is used to indicate whether it is true positive detection or background. We model the object relations by considering the relative depth ordering between them. To this end, we divide the proposals into the singleton object set $\mathcal{S}$, and the overlapping object set $\mathcal{P}$ which consists of occluding object pairs. For each $p \in \mathcal{P}$ and $p = (m,n)$, we introduce h_{mn} to describe their occlusion relations. Denoting the supervoxel labeling, object states and their relations of the entire sequence as $\mathbf{L}$, $\mathbf{D}$, $\mathbf{H}$, we define the overall energy function of our CRF model as follows:

$$E(\mathbf{L},\mathbf{D},\mathbf{H}|\mathcal{T}) = E_v(\mathbf{L}) + \sum_{m\in\mathcal{S}} E_s^m(\mathbf{L},d_m) + \sum_{p\in\mathcal{P}} E_r^p(\mathbf{L},d_m,d_n,h_{mn}) \quad (1)$$

where E_v denotes the potentials at supervoxel level. E_s^m and E_r^p are potentials for singleton and overlapped objects.

We first develop an efficient mean field inference algorithm to jointly infer the supervoxel labels, object activations and their relations for a moderate number of object proposals. For scaling up inference with many object proposals, we propose to select an informative subset of objects and their relation nodes [6]. To this end, we build a set of subgraphs corresponding to the object hypotheses, which are selected in our inference procedure. Specifically, we introduce a subgraph selection state vector $\mathbf{z} = [\mathbf{z}^s, \mathbf{z}^r]$, where $\mathbf{z}^s$ and $\mathbf{z}^r$ are for singleton and object pair set $\mathcal{S},\mathcal{P}$ respectively. Each element z_k in $\mathbf{z}$ is an binary indicator and $z_k = 1$ means subgraph k is selected. The full CRF model with subgraph selection can be defined by its energy function, $E(\mathbf{L},\mathbf{D},\mathbf{H},\mathbf{z}|\mathcal{T}) = E_v + \sum_{m\in\mathcal{S}} z_m^s E_s^m + \sum_{p\in\mathcal{P}} z_p^r E_r^p$.

We formulate the subgraph selection as a Markov Decision Process (MDP) and develop a learning approach to search the optimal policy for sequentially choosing most informative subgraphs. We define a reward function using the improvement on average per-class pixel accuracy, and learn an approximate policy based on Q-learning [3]. Our policy takes long-range features generated by both current model uncertainty and video input, and predicts the most valuable subgraph to choose in next step. Furthermore, we also use an imitation learning scheme [1] to train a fast local classifier that approximates the optimal decision.

We evaluate our approach on three publicly available semantic video segmentation datasets. We demonstrate that our learned policy is capable of selecting informative object proposals and their relations, leading to much simpler model structure and comparable and even higher segmentation accuracy. Figure 2 shows the prediction curves for average class accuracy and foreground object accuracy on CamVid Dataset [2]. Our proposed methods (first three in legend) achieve a better traded-off in accuracy and efficiency than baseline methods.

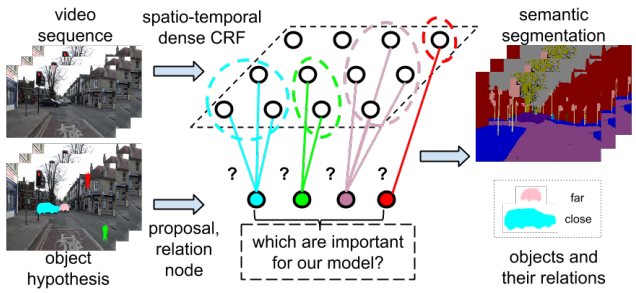

Figure 1: Overview of our approach. Example of the object-augmented dense CRF model. Our active inference adaptively selecting subgraphs thus improve the inference efficiency.

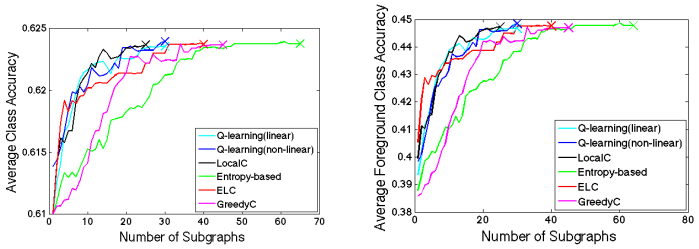

Figure 2: Traded-off performance on the CamVid dataset. The curve shows the increase in accuracy over the selective inference model as a function of subgraph number. The cross shows the termination point for inference.

Table 1 shows the comparison of the efficiency of our algorithm and two state-of-the-art methods. Our algorithm is much faster, and its inference time increases only sub-linearly with respect to the number of hypotheses.

# of Subgraphs	Time(s)		
	[4] (GraphCut)	FullCRF (MeanField)	Our Method
21.6	4.3	2.6	**1.5**
41.1	5.8	3.3	**1.6**
81.8	8.3	5.3	**1.8**
165	14.4	10.9	**2.4**

Table 1: Inference efficiency v.s. number of proposals of different methods on two-second video chunks in CamVid.

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

[2] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008.

[3] Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003.

[4] Buyu Liu, Xuming He, and S. Gould. Multi-class semantic video segmentation with exemplar-based object reasoning. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, 2015.

[5] Joseph Tighe and Svetlana Lazebnik Marc Niethammer. Scene parsing with object instances and occlusion ordering. In *CVPR*, 2014.

[6] DJ Weiss and Ben Taskar. Learning Adaptive Value of Information for Structured Prediction. In *Advances in Neural Information Processing (NIPS)*, 2013.

[7] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.