Classifier Based Graph Construction for Video Segmentation

Anna Khoreva[1], Fabio Galasso[2], Matthias Hein[3], Bernt Schiele[1]
[1]Max Planck Institute for Informatics, Germany. [2]OSRAM Corporate Technology, Germany. [3]Saarland University, Germany.

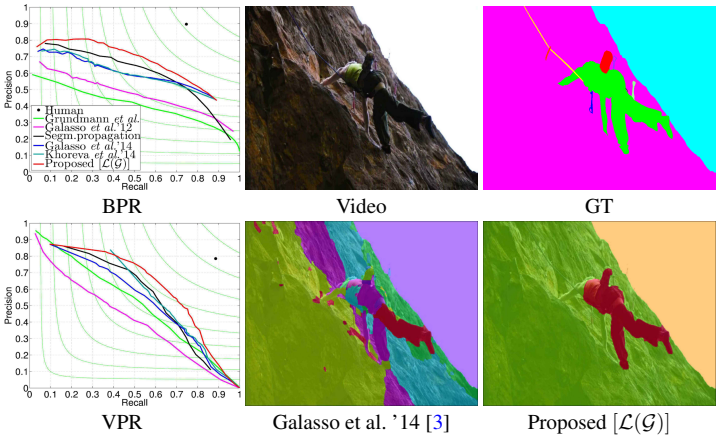
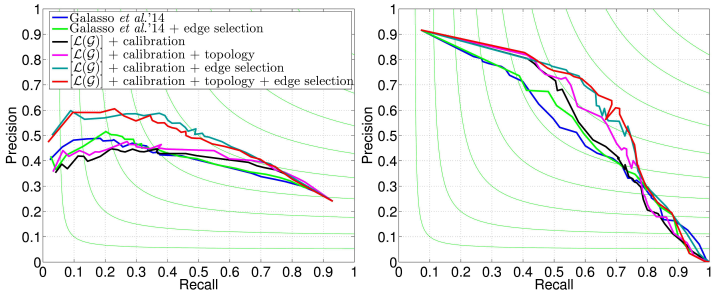


Figure 1: Climbing up! We contribute theory and best-practices for *graph construction*, an essential part of video segmentation pipelines. In combination with state-of-the-art features and a partitioning model [3], our proposed algorithm sets a new state-of-the-art performance on the challenging VSB100 [2].



	BPR			VPR		
Algorithm	ODS	OSS	AP	ODS	OSS	AP
Galasso *et al.* CVPR'14 [3]	0.46	0.49	0.37	0.54	0.61	0.56
Galasso *et al.* CVPR'14 [3] + edge selection	0.47	0.51	0.37	0.57	0.62	0.57
$[\mathcal{L}(\mathcal{G})]$ + calibration	0.49	0.53	0.36	0.59	0.65	0.59
$[\mathcal{L}(\mathcal{G})]$ + calibration + topology	0.51	0.54	0.38	0.62	0.67	0.62
$[\mathcal{L}(\mathcal{G})]$ + calibration + edge selection	**0.52**	0.57	**0.44**	0.65	**0.70**	0.64
$[\mathcal{L}(\mathcal{G})]$ + calibration + topology + edge selection	**0.52**	**0.58**	**0.44**	**0.66**	**0.70**	**0.65**

Figure 2: Comparison of the proposed graph learning method $[\mathcal{L}(\mathcal{G})]$ with the baseline algorithm of [3], on the validation set of VSB100 [2].

Video segmentation has become an important and active research area. Graph-based approaches are among the top-performing methods for video segmentation. Graph-based video segmentation techniques consist of three essential components: 1. **features extraction**: powerful features among pairs of pixels or superpixels account for object appearance and motion similarities; 2. **graph construction**: spatio-temporal neighborhoods of pixels or superpixels (the graph edges) are modeled using a combination of those features; 3. **graph partitioning**: video segmentation is formulated as a graph partitioning problem.

We address in this work graph construction, which has so far received little attention. We propose and empirically evaluate procedures to learn both the edge topology and weights.

Proposed graph construction

We build upon the graph partitioning model of [3] and show that addressing the graph construction explicitly helps to achieve better performance (cf. Fig. 1) without altering the graph partitioning or the underlying features.

Adopting learning allows to seamlessly integrate an arbitrary number of features into the computation of the graph edge weights, letting the Random Forest classifier work out the optimal combination. There are four superpixel edge types: within, across 1, across 2 and across >2 frames. We set therefore to consider four classifiers for the four edge types and for each type we validate the subset of features to improve the model.

In the case of connectivity between superpixels within or across 1 frame, the graph of [3] is densified by using edges among neighboring superpixels (*layer-1 neighbors*) and among more distant superpixels which share the same neighbor (*layer-2 neighbors*). We propose to treat the topologically different neighbors separately. We separate the two topologies both for the within and the across 1 type and re-learn separate classifiers. The across 2 and across >2 types only have layer-1 neighbors and is therefore not affected by the topological procedure. Treating separately the two layers helps to increase video segmentation performance (cf. Fig. 2).

An ideal subsequent processing of the graph would be the selection of the most likely edges and the deletion of wrong ones. This is desirable as it sparsifies the graph and reduces the chance of segmentation errors. We propose a probabilistic interpretation of the learnt scores and calibration of the classifier outputs based on their performance on the validation set. For each affinity type we define a linear mapping $\Pi : S \mapsto P$, such that the classifier score s is approximated by its precision value p. This calibration serves to align the classifiers scores to their quality and is important when combining multiple classifiers. The calibrated classifier scores are used as edge weights in the graph.

We then modify the graph structure by selecting the edges with high confidence. Each affinity type is thresholded with some confidence level, reducing the number of edges in the graph. The goal is to have a connected graph with a minimal set of the most certain edges, as for maximal sparsity and the least chance of segmentation error. For finding the optimal thresholds for each affinity type a grid search is applied.

Evaluation

In Figure 2, we analyze how the learnt graph $[\mathcal{L}(\mathcal{G})]$ and the proposed steps improve performance on the validation set of VSB100 [2], with respect to the baseline [3]. Given a learnt and calibrated graph, topology improves 2.2% while edge selection improves by 5.2%. Edge selection contributes therefore more than topology. To further test the importance of edge selection, we have applied this to the baseline [3]. The improvement is only marginal - 1.3%. We conclude that a pre-requisite for the successful edge selection is weight calibration plus the good performance of the classifier.

We further compare the proposed method $[\mathcal{L}(\mathcal{G})]$ to the state-of-the-art video segmentation algorithms [1, 2, 3, 4, 5] on the test set of VSB100 [2] (cf. Fig. 1). By learning the graph, we improve the results of the best performing algorithm by 6%, while reducing its runtime by 55%, as the learnt graph is much sparser (the average number of edges is reduced to 15%).

[1] F. Galasso, R. Cipolla, and B. Schiele. Video segmentation with superpixels. In *ACCV*, 2012.
[2] F. Galasso, N. S. Nagaraja, T. Z. Cardenas, T. Brox, and B. Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *ICCV*, 2013.
[3] F. Galasso, M. Keuper, T. Brox, and B. Schiele. Spectral graph reduction for efficient image and streaming video segmentation. In *CVPR*, 2014.
[4] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010.
[5] A. Khoreva, F. Galasso, M. Hein, and B. Schiele. Learning must-link constraints for video segmentation based on spectral clustering. In *GCPR*, 2014.

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