A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis

Fabio Galasso[1], Naveen Shankar Nagaraja[2], Tatiana Jiménez Cárdenas[2], Thomas Brox[2], Bernt Schiele[1]
[1] Max Planck Institute for Informatics, Germany
[2] University of Freiburg, Germany

Abstract

Video segmentation research is currently limited by the lack of a benchmark dataset that covers the large variety of subproblems appearing in video segmentation and that is large enough to avoid overfitting. Consequently, there is little analysis of video segmentation which generalizes across subtasks, and it is not yet clear which and how video segmentation should leverage the information from the still-frames, as previously studied in image segmentation, alongside video specific information, such as temporal volume, motion and occlusion. In this work we provide such an analysis based on annotations of a large video dataset, where each video is manually segmented by multiple persons. Moreover, we introduce a new volume-based metric that includes the important aspect of temporal consistency, that can deal with segmentation hierarchies, and that reflects the tradeoff between over-segmentation and segmentation accuracy.

1. Introduction

Video segmentation is a fundamental problem with many applications such as action recognition, 3D reconstruction, classification, or video indexing. Many interesting and successful approaches have been proposed. While there are standard benchmark datasets for still image segmentation, such as the Berkeley segmentation dataset (BSDS) [18], a similar standard is missing for video segmentation. Recent influential works have introduced video datasets that specialize on subproblems in video segmentation, such as motion segmentation [4], occlusion boundaries [22, 23], or video superpixels [28].

This work aims for a dataset with corresponding annotation and an evaluation metric that can generalize over subproblems and help in analyzing the various challenges of video segmentation. The proposed dataset uses the natural sequences from [23]. In contrast to [23], where only a single frame of each video is segmented by a single person, we extend this segmentation to multiple frames and multiple persons per frame. This enables two important properties in the evaluation metric: (1) temporally inconsistent segmentations are penalized by the metric; (2) the metric can take the ambiguity of correct segmentations into account.

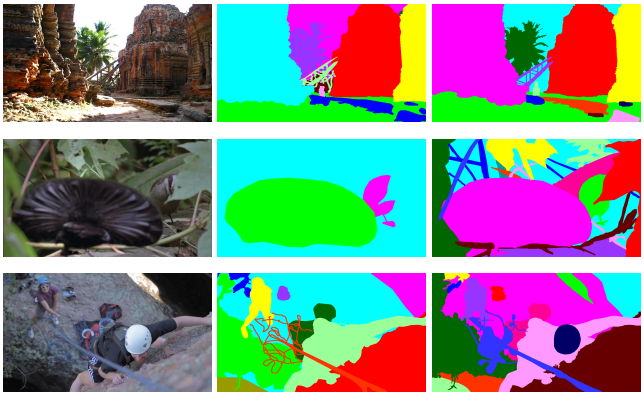


Figure 1. (Column-wise) Frames from three video sequences of the dataset [23] and 2 of the human annotations which we collected for each. Besides the different coloring, the frames show different levels of human agreement in labelling. We provide an analysis of state-of-the-art video segmentation algorithms by means of novel metrics leveraging multiple groundtruths. We also analyze additional video specific subproblems, such as motion, non-rigid motion and camera motion.

The latter property has been a strong point of the BSDS benchmark on single image segmentation [18]. Some segmentation ambiguities vanish by the use of videos (e.g., exact placement of a boundary due to similar color and texture), but the scale ambiguity between scene elements, objects and their parts persists. Should a head be a separate segment or should the person be captured as a whole? Should a crowd of people be captured as a whole or should each person be separated? As in [18], we approach the scale issue by multiple human annotations to measure the natural level of ambiguity and a precision-recall metric that allows to compare segmentations tuned for different scales.

Moreover, the dataset is supposed to cover also the various subproblems of video segmentation. To this end, we provide additional annotation that allows an evaluation exclusively for moving/static objects, rigid/non-rigid objects or videos taken with a moving/static camera. The annotation also enables deeper analysis of typical limitations of

current video segmentation methods. We consider a large set of publicly available methods and compare their performance on the general benchmark and the different subsets.

2. Video Segmentation Literature

A large body of literature exists on video segmentation leveraging appearance [2, 27, 13, 29], motion [21, 4, 11], or multiple cues [8, 12, 14, 15, 20, 16, 17, 19, 10, 6, 9].

Various techniques are used, e.g. generative layered models [14, 15], graph-based models [13], mean-shift [8, 20] and techniques based on manifold-embedding and eigendecomposition such as ISOMAP [11] and spectral clustering [21, 4, 9, 10]. Graph-based [13, 29] and mean-shift techniques [20] are based on local properties and usually focus on generating an over-segmentation of a video. Other methods are more object centric. For instance, layered models [14, 15] have shown potential to learn object motion and appearance. Recent works on video segmentation exploit the motion history contained in point trajectories [4, 17, 19, 9]. These methods focus only on moving objects, while static objects are combined to a single cluster. Instead of points, some other works [2, 27, 11, 10, 6] track superpixels as these provide a desirable computational reduction and powerful within-frame representation.

This literature overview, which is by far not complete, shows the large diversity of video segmentation approaches. There is a fundamental need for a common dataset and benchmark evaluation metrics that can cover all the various subtasks and that allows an analysis highlighting the strengths and limitations of each approach. In the following, we will first specify criteria for the dataset and annotation in Section 3; then we specify the evaluation metric in Section 4, and finally we will use these to analyze the current state of the field in Sections 5 and 6.

3. Video Segmentation Dataset and Annotation

A good video segmentation dataset should consist of a large number of diverse sequences with the diversity spanning across different aspects. Some of those aspects are equivalent to those of a good image segmentation dataset, i.e., the videos should contain a variety of "objects": people, animals, man-made and natural scene elements; the size of the objects should span a wide range: from few hundred pixels to covering a large portion of the frame; the appearance of the scene elements should vary across sequences and should comprise of both homogenous and textured objects. In addition to the image based diversity, the diversity of video sequences should also include occlusion and different kinds of object and camera motion: translational, scaling and perspective motion.

Current video segmentation datasets are limited in the aforementioned aspects: figment [9] only includes equally-sized basketball players; CamVid [3] fulfills appearance heterogeneity but only includes 4 sequences, all of them recorded from a driving car. Similarly [5, 25, 13, 28] only include few sequences and [13] even lacks annotation. On motion segmentation, the Hopkins155 dataset [24] provides many sequences, but most of them show artificial checkerboard patterns and ground truth is only available for a very sparse set of points. The dataset in [4] offers dense ground truth for 26 sequences, but the objects are mainly limited to people and cars. Moreover, the motion is notably translational. Additionally, these datasets have at most VGA quality (only [3] is HD) and none of them provide viable training sets.

A recent video dataset, introduced in [23] for occlusion boundary detection, fulfills the desired criteria of diversity. While the number of frames per video is limited to a maximum of 121 frames, the video sequences are HD and include 100 videos arranged into 40 train + 60 test sequences. The dataset is also challenging for current video segmentation algorithms as experiments show in Sections 5 and 6. We adopt this dataset and provide the annotation necessary to make it a general video segmentation benchmark.

Following [18] we provide multiple human annotations per frame. Video annotations should be accurate at object boundaries and most importantly temporally consistent: an object should have the same label in all ground truth frames throughout the video sequence. To this end, we invited the annotators to watch the videos completely and thus motion played a major role in their perception. Then, they were given the following pointers: *What to label? The objects, e.g. people, animals etc. and image parts, e.g. water surfaces, mountains, which describe the video sequence.*

4. Benchmark evaluation metrics

We propose to benchmark video segmentation performance with a boundary oriented metric and with a volumetric one. Both metrics make use of M human segmentations and both can evaluate over- and under-segmentations.

4.1. Boundary precision-recall (BPR)

The boundary metric is most popular in the BSDS benchmark for image segmentation [18, 1]. It casts the boundary detection problem as one of classifying boundary from non-boundary pixels and measures the quality of a segmentation boundary map in the precision-recall framework:

$$P = \frac{|S \cap \left(\bigcup_{i=1}^{M} G_i\right)|}{|S|} \quad (1)$$

$$R = \frac{\sum_{i=1}^{M} |S \cap G_i|}{\sum_{i=1}^{M} |G_i|} \quad (2)$$

$$F = \frac{2PR}{R+P} \quad (3)$$

where S is the set of machine generated segmentation boundaries and $\{G_i\}_{i=1}^{M}$ are the M sets of human annotation boundaries. The so-called F-measure is used to evaluate aggregate performance. The intersection operator $\cap$ solves a bipartite graph assignment between the two boundary maps.

The metric is of limited use in a video segmentation benchmark, as it evaluates every frame independently, i.e., temporal consistency of the segmentation does not play a role. Moreover, good boundaries are only half the way to a good segmentation, as it is still hard to obtain closed object regions from a boundary map. We keep this metric from image segmentation, as it is a good measure for the localization accuracy of segmentation boundaries. The more important metric, though, is the following volumetric metric.

4.2. Volume precision-recall (VPR)

VPR optimally assigns spatio-temporal volumes between the computer generated segmentation $\mathbb{S}$ and the M human annotated segmentations $\{\mathbb{G}_i\}_{i=1}^{M}$ and measures their overlap. A preliminary formulation that, as we will see, has some problems is

$$\tilde{P} = \frac{1}{M}\sum_{i=1}^{M}\frac{\sum_{s\in\mathbb{S}}\max_{g\in\mathbb{G}_i}|s\cap g|}{|\mathbb{S}|} \quad (4)$$

$$\tilde{R} = \sum_{i=1}^{M}\frac{\sum_{g\in\mathbb{G}_i}\max_{s\in\mathbb{S}}|s\cap g|}{\sum_{i=1}^{M}|\mathbb{G}_i|} \quad (5)$$

The volume overlap is expressed by the intersection operator $\cap$ and $|.|$ denotes the number of pixels in the volume. A maximum precision is achieved with volumes that do not overlap with multiple ground truth volumes. This is relatively easy to achieve with an over-segmentation but hard with a small set of volumes. Conversely, recall counts how many pixels of the ground truth volume are explained by the volume with maximum overlap. Perfect recall is achieved with volumes that fully cover the human volumes. This is trivially possible with a single volume for the whole video.

Obviously, degenerate segmentations (one volume covering the whole video or every pixel being a separate volume) achieve relatively high scores with this metric. The problem can be addressed by a proper normalization, where the theoretical lower bounds (achieved by the degenerate segmentations) are subtracted from the overlap score:

$$P = \frac{\sum_{i=1}^{M}\Big[\{\sum_{s\in\mathbb{S}}\max_{g\in\mathbb{G}_i}|s\cap g|\} - \max_{g\in\mathbb{G}_i}|g|\Big]}{M|\mathbb{S}| - \sum_{i=1}^{M}\max_{g\in\mathbb{G}_i}|g|} \quad (6)$$

$$R = \frac{\sum_{i=1}^{M}\sum_{g\in\mathbb{G}_i}\{\max_{s\in\mathbb{S}}|s\cap g| - 1\}}{\sum_{i=1}^{M}\{|\mathbb{G}_i| - \Gamma_{\mathbb{G}_i}\}} \quad (7)$$

where $\Gamma_{\mathbb{G}_i}$ is the number of ground truth volumes in $\mathbb{G}_i$.

For both BPR and VPR we report average precision (AP), the area under the PR curve, and optimal aggregate measures by means of the F-measures: optimal dataset scale (ODS), aggregated at a fixed scale over the dataset, and optimal segmentation scale (OSS), optimally selected for each segmentation. In the case of VPR, the F-measure coincides with the Dice coefficient between the assigned volumes.

4.3. Properties of the proposed metrics

BPR and VPR satisfy important requirements (cf. [26]):
i. Non-degeneracy: measures are low for degenerate segmentations.
ii. No assumption about data generation: the metrics do not assume a certain number of labels and apply therefore to cases where the computed number of labels is different from the ground truth.
iii. Multiple human annotations: inconsistency among humans to decide on the number of labels is integrated into the metrics, which provides a sample of the acceptable variability.
iv. Adaptive accommodation of refinement: segmentation outputs addressing different coarse-to-fine granularity are not penalized, especially if the refinement is reflected in the human annotations, but granularity levels closer to human annotations score higher than the respective over- and under-segmentations; this property draws directly from the humans, who psychologically perceive the same scenes to a different level of detail.
v. Coarse-to-fine segmentations and working regimes: the metrics allow the insightful analysis of algorithms at different working regimes: over-segmentation algorithms decomposing the video into several smaller temporally consistent volumes will be found in the high precision VPR area and correspondingly in the high recall BPR area. More object-centric segmentation methods that tend to yield few larger object volumes will be found in the VPR high recall area, BPR high precision area. In both regimes, algorithms that trade off precision and recall in a slightly different manner can be compared in a fair way via the F-measure.
vi. Compatible scores: both metrics allow comparing the results of the same algorithm on different videos and the results of different algorithms on the same set of videos.

The VPR metric additionally satisfies the requirement of **vii. Temporal consistency**: object labels that are not consistent over time get penalized by the metric.

All previously proposed metrics do not satisfy all these constraints. The one in [4] is restricted to motion segmentation and does not satisfy (iii) and (v). The metrics in [28] do not satisfy (iii). The boundary metric in [1] is designed for still image segmentation and do not satisfy (vii). The region metrics in [1] have been extended to volumes [27, 10] but do not satisfy (i) and (v).

	BPR			VPR			Length	NCL
Algorithm	ODS	OSS	AP	ODS	OSS	AP	$\mu(\delta)$	μ
Human	0.71	0.71	0.53	0.83	0.83	0.70	83.24(40.04)	11.90
*Corso et al. [7]	0.51	0.53	0.37	0.51	0.52	0.38	70.67(48.39)	25.83
*Galasso et al. [10]	**0.52**	**0.56**	**0.44**	0.45	0.51	0.42	80.17(37.56)	8.00
*Grundmann et al. [13]	0.47	0.54	0.42	**0.52**	**0.55**	**0.52**	87.69(34.02)	18.83
*Ochs and Brox [19]	0.14	0.14	0.04	0.25	0.25	0.12	87.85(38.83)	3.73
Xu et al. [29]	0.40	0.48	0.33	0.45	0.48	0.44	59.27(47.76)	26.58
IS - Arbelaez et al. [1]	0.61	0.65	0.61	0.26	0.27	0.16	1.00(0.00)	4389.74
Baseline	0.60	0.64	0.57	0.59	0.62	0.56	25.50(36.48)	258.05
Oracle & IS - Arbelaez et al. [1]	0.61	0.67	0.61	0.65	0.67	0.68	-	118.56

Table 1. Aggregate performance evaluation of boundary precision-recall (BPR) and volume precision-recall (VPR) of state-of-the-art VS algorithms. We report optimal dataset scale (ODS) and optimal segmentation scale (OSS), achieved in term of F-measure, alongside the average precision (AP), e.g. area under the PR curve. Corresponding mean (μ) and standard deviation statistics (δ) are shown for the volume lengths (Length) and the number of clusters (NCL). (*) indicates evaluated on video frames resized by 0.5 in the spatial dimension, due to large computational demands.

5. Evaluation of Segmentation Algorithms

This and the following section analyze a set of state-of-the-art video segmentation algorithms on the general problem and in scenarios which previous literature has addressed with specific metrics and datasets: supervoxel segmentation, object segmentation and motion segmentation.

5.1. Variation among human annotations

The availability of ground-truth from multiple annotators allows the evaluation of each annotator's labeling against others'. The human performance from Table 1 and Figure 2 expresses the difficulty of the dataset. In particular, BPR plots indicate high precision, which reflects the very strong human capability to localize object boundaries in video material. Recall is lower, as different annotators label scenes at different levels of detail [18, 26]. As expected, humans reach high performance with respect to VPR, which shows their ability to identify and group objects consistently in time. Surprisingly, human performance slightly drops for the subset of moving objects, indicating that object knowledge is stronger than motion cues in adults.

5.2. Selection of methods

We selected a number of recent state-of-the-art video segmentation algorithms based on the availability of public code. Moreover, we aimed to cover a large set of different working regimes: [19] provides a *single* segmentation result and specifically addresses the estimation of the number of moving objects and their segmentation. Others [7, 13, 10, 29] provide a *hierarchy* of segmentations and therefore cover multiple working regimes. According to these working regimes, we separately discuss the performance of the methods in the (VPR) high-precision area (corresponding to super-voxelization) and the (VPR) high-recall area (corresponding to object segmentation with a tendency to under-segmentation).

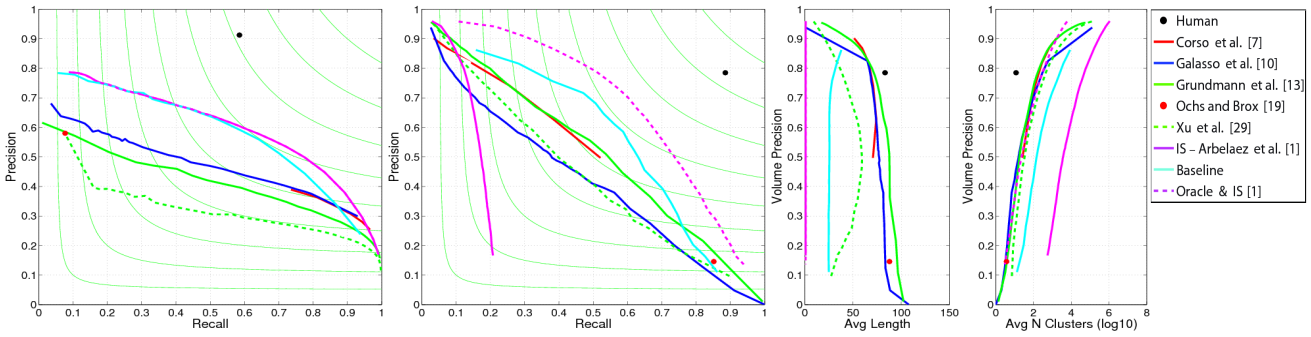
5.3. Supervoxelization

Several algorithms [7, 13, 29] have been proposed to *over*-segment the video as a basis for further processing. [10] provide coarse-to-fine video segmentation and could be additionally employed for the task. [28] defined important properties for the supervoxel methods: supervoxels should respect object boundaries, be aligned with objects without spanning multiple of them (known as *leaking*), be temporally consistent and parsimonious, i.e. the fewer the better.

An ideal supervoxelization algorithm preserves all boundaries at the cost of over-segmenting the video. In the proposed PR framework this corresponds to the high-recall regime of the BPR curve, in Figure 2. BPR takes into account multiple human annotations, i.e. perfect recall is achieved by algorithms detecting all the boundaries identified by all human annotators. Algorithms based on spectral clustering [7, 10] slightly outperform the others, providing supervoxels with more compact and homogeneous shapes, as also illustrated by the sample results in Figure 3.

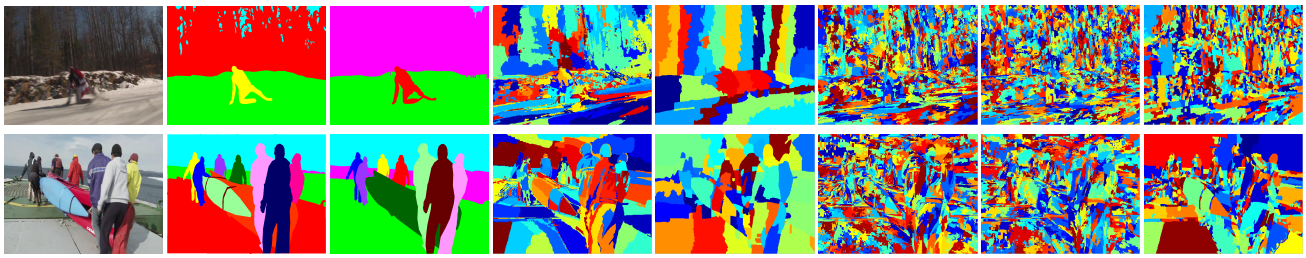
Temporal consistency and leaking are benchmarked by VPR in the high-precision area. VPR, like BPR, is also consistent with the multiple human annotators: perfect volume precision is obtained by supervoxels not leaking any of the multiple GT's. Greedy-merge graph-based methods [13, 29] prove better supervoxelization properties, as the irregular supervoxel shapes better adapt to the visual objects.

Statistics on the supervoxel lengths and on their number complete the proposed supervoxel evaluation. In Figure 2, the volume precision is plotted against the average volume length and their cardinality: this illustrates how much a segmentation algorithm needs to over-segment the video (number of clusters (NCL)) to achieve a given level of precision, and what the average length of volumes is at that precision level. [13, 29, 10] all recur to more than 50000 clusters to achieve best precision, but [13] also maintains volumes of $\sim$20 frames, as opposed to $\sim$9 frames for [29].



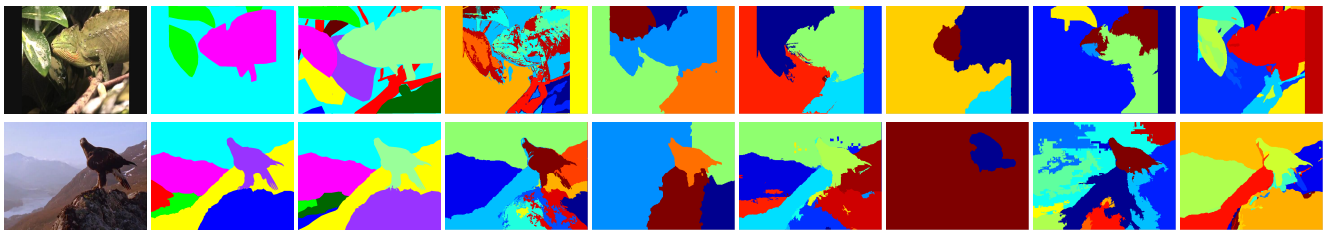
BPR VPR Length NCL

Figure 2. Boundary precision-recall (BPR) and volume precision-recall (VPR) curves benchmark state-of-the-art VS algorithms at different working regimes, from over-segmentation (high-recall BPR, high-precision VPR) to object-centric segmentations providing few labels for the video (high-precision BPR, high-recall VPR). Mean length - volume precision and mean number of clusters - volume precision curves (respectively Length and NCL) complement the PR curves. These provide insights into how different algorithms fragment differently the spatial and temporal domains, i.e. the number of volumes and their length.



Video GT-1 GT-2 [7] [10] [13] [29] Baseline

Figure 3. Sample supervoxel results provided by algorithms [7, 13, 29, 10].



Video GT-1 GT-2 [7] [10] [13] [19] [29] Baseline

Figure 4. Sample results provided by the selected VS algorithms for object segmentation.

5.4. Object segmentation

Object segmentation stands for algorithms identifying the visual objects in the video sequence and reducing the over-segmentation. [4, 19] have targeted this particular subproblem in the case of motion segmentation.

Intuitively, important aspects are the parsimonious detection of salient object boundaries and the segmentation of the video sequences into volumes which "explain" (i.e. cover, overlap) entire objects or groups of them, maintaining temporal consistency. Each of the visual objects should be covered by a single volume over the entire video, at the cost of having volumes overlapping multiple objects.

An ideal object segmentation method detects the few salient boundaries in the video. This corresponds to the high-precision regime of the BPR curve in Figure 2. Methods providing the boundaries on which *all* human annotators agree achieve highest precision.

The volume property to explain visual objects maintaining temporal consistency is benchmarked by VPR in the high-recall regimes. All algorithms may achieve perfect recall by labeling the whole video as one object. As for the normalization of VPR (Equations 6,7), the degenerate output achieves 0 precision. We propose therefore to compare the selected algorithms at a minimum level of $\sim 15\%$ precision.

Statistics on the average length (Length) and number of clusters (NCL) at the given $\sim 15\%$ volume precision in Figure 2 complement the object segmentation analysis. In particular all algorithms qualifying for the task [13, 19, 29, 10]

	Motion segmentation						Non-rigid motion						Moving camera					
	BPR			VPR			BPR			VPR			BPR			VPR		
Algorithm	ODS	OSS	AP	ODS	OSS	AP	ODS	OSS	AP	ODS	OSS	AP	ODS	OSS	AP	ODS	OSS	AP
Human	0.59	0.59	0.36	0.76	0.76	0.59	0.52	0.52	0.29	0.77	0.77	0.60	0.72	0.72	0.54	0.81	0.81	0.66
*Corso et al. [7]	0.21	0.21	0.12	0.37	0.35	0.23	0.18	0.18	0.10	**0.34**	0.33	0.21	**0.53**	0.54	0.38	0.51	0.52	0.38
*Galasso et al. [10]	**0.34**	**0.43**	**0.23**	**0.42**	**0.46**	**0.36**	**0.29**	**0.40**	**0.18**	0.30	0.33	0.23	0.52	**0.57**	**0.45**	0.44	0.49	0.40
*Grundmann et al. [13]	0.25	0.34	0.15	0.37	0.41	0.32	0.24	0.31	0.12	0.33	**0.38**	**0.28**	0.49	0.56	0.42	**0.53**	**0.55**	**0.52**
*Ochs and Brox [19]	0.26	0.26	0.08	0.41	0.41	0.23	0.18	0.18	0.05	0.17	0.17	0.08	0.14	0.14	0.04	0.24	0.24	0.11
Xu et al. [29]	0.22	0.33	0.16	0.32	0.36	0.27	0.19	0.28	0.13	0.29	0.32	0.22	0.41	0.49	0.34	0.45	0.48	0.43
IS - Arbelaez et al. [1]	0.46	0.36	0.35	0.22	0.22	0.13	0.40	0.33	0.27	0.19	0.19	0.11	0.61	0.66	0.60	0.26	0.27	0.17
Baseline	0.45	0.50	0.33	0.52	0.57	0.47	0.40	0.48	0.25	0.49	0.55	0.41	0.60	0.65	0.56	0.57	0.60	0.53
Oracle & IS [1]	0.46	0.36	0.35	0.59	0.60	0.60	0.40	0.34	0.27	0.58	0.60	0.58	0.61	0.68	0.60	0.65	0.67	0.69

Table 2. Aggregate measures of boundary precision-recall (BPR) and volume precision-recall (VPR) for the VS algorithms on the motion subtasks. (*) indicates evaluated on video frames resized by 0.5 in the spatial dimension, due to large computational demands.

provide few clusters (less than 10), but only [13, 19, 10] provide volumes lasting more than 80 frames.

Figure 4 illustrates best outputs of the selected algorithms on object segmentation. Some [7, 13, 29] consistently tend to oversegment the scene, while others [19, 10] may over-simplify it, missing some visual objects.

5.5. Image segmentation and a baseline

We also have benchmarked a state-of-the-art image segmentation algorithm [1], alongside an associated oracle performance and a proposed baseline. [1] provides only a per-frame segmentation. In the proposed framework, it can be directly compared to video segmentation methods with regard to the boundary BPR metric, but it is heavily penalized on the VPR metric, where temporal consistency across frames is measured. These characteristics are observable in Figure 2 and Table 1: average length (Length) reads 1 and number of clusters (NCL) is larger w.r.t. video segmentation algorithms by two orders of magnitude, as image segments re-initialize at each frame of the video sequences, $\sim$100 frame long.

Interestingly, [1] consistently outperforms all selected video segmentation algorithms on the boundary metric, indicating that current video segmentation methods are not proficient in finding good boundaries. To test performance that [1] would achieve on the VPR metric, if it addressed temporal consistency perfectly, we resort to an oracle. The dashed magenta curve in the VPR illustrates performance of [1], when the per-frame segmentation is bound into volumes over time by an oracle prediction based on the ground truth. The performance on the volume reaches levels far beyond state-of-the-art video segmentation performance.

This motivates to introduce a new baseline (cyan curve). The result of [1] (across the hierarchy) at the central frame of the video sequences are propagated to other frames in the video with optical flow [30] and used to label corresponding image segments (across the hierarchy) by maximum voting. As expected, the working regimes of the simple baseline do not extend to superpixelization nor to segmenting few objects, due to the non-repeatability of image segmentation (cf. [11]). This is more pronounced at hierarchical levels farther from the scale of the visual objects, as both large image segments and fine superpixels are in many images arbitrary groupings, likely to change over time due to lack of temporal consistency. Although simple, this baseline outperforms all considered video segmentation algorithms [7, 13, 19, 29, 10], consistently over the mid-recall range (however with low average length and large number of clusters, undesirable quality of a real video segmenter).

This analysis suggests that state-of-the-art image segmentation is at a more mature research level than video segmentation. In fact, video segmentation has the potential to benefit from additional important cues, e.g. motion, but models are still limited by computational complexity.

5.6. Computational complexity

Only the streaming algorithm of [29] could be tested on the full HD quality videos. [19] follows [29] in the ranking, as it leverages on sparse tracks and per-frame densification. The agglomerative algorithms were computationally demanding in terms of memory and time, while [10] was the most costly, due to the global spectral partitioning.

6. Evaluation of Motion Segmentation Tasks

Compared to still images, videos provide additional information that add to the segmentation into visual objects. Motion is certainly the most prominent among these factors. Motion comes with several nuances, e.g. object vs camera motion, translational vs zooming and rotating, and it affects a number of other video specificities, e.g. temporal resolution leading to slow vs. fast motion, motion blur, partial- and self-occlusions, objects leaving and appearing in the scene.

The additional indication of moving objects allows us to

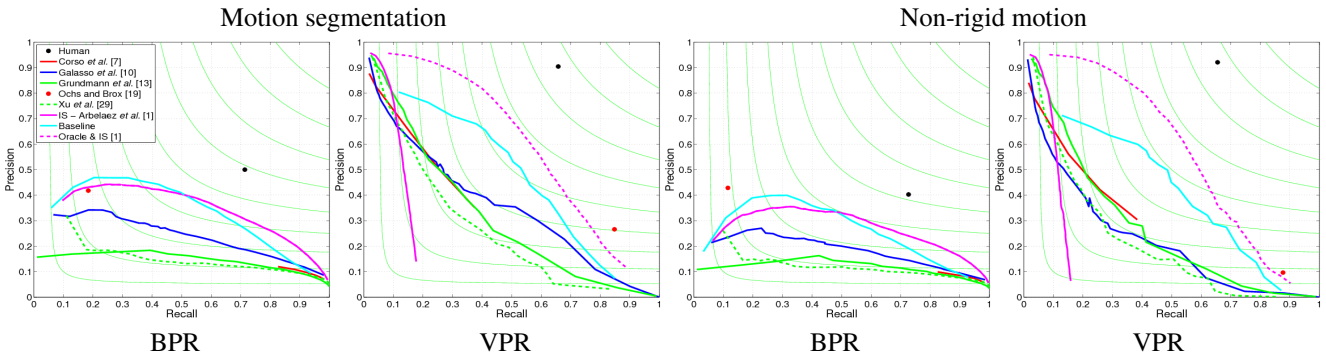


Figure 5. Boundary precision-recall (BPR) and volume precision-recall (VPR) curves for VS algorithms on two motion subtasks.

Figure 6. Sample results provided by the selected VS algorithms for the task of motion segmentation.

compare the performance of segmentation methods on moving objects vs all objects. On one hand, motion comes with additional problems, such as occlusions, change of lighting, etc.; on the other hand, motion helps segmenting objects as a whole, which is generally impossible in still images.

We investigated how well the discussed methods exploited the motion cue. For the first two graphs in Figure 5 and the statistics in Table 2 (first block-column), all non-moving objects were ignored in the evaluation. As expected, the method from [19], which focuses on motion segmentation, performs much better than on the general benchmark. Especially precision goes up, as the method is no longer penalized for combining static objects to a single volume. Most other video segmentation methods perform worse on the moving objects than on the static ones. It should be noted how the similar VPR aggregate F-measures in Table 2 for [10] and [19] correspond to different segmentation qualities: [10] achieves better boundary precision but it over-segments the moving objects; by contrast [19] identifies the moving objects with fewer less-boundary-aligned volumes. This may also be observed from the sample results for the subtask in Figure 6.

Among the moving objects, we separately investigated objects undergoing non-rigid motion. As from the second two graphs in Figure 5 and the corresponding statistics in Table 2, the performance went down for all methods and also hit the best motion subtask performers. There is much room for improvement.

The performance of the still image segmentation algorithm of [1] and the proposed baseline is also interesting. While [1] is strongly penalized by VPR for the missing temporal consistency, it outperforms all considered video segmentation algorithms on the boundary metric, where the

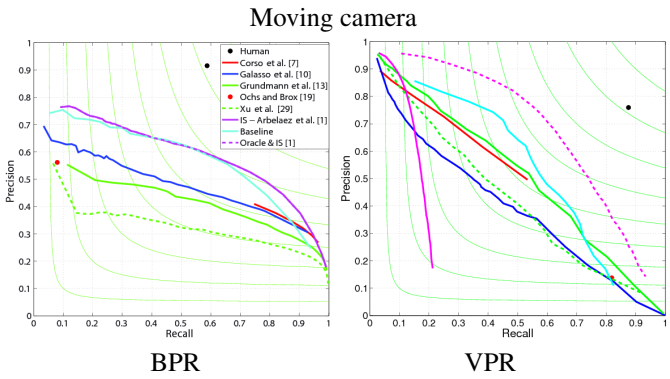


Figure 7. Boundary precision-recall (BPR) and volume precision-recall (VPR) curves for the selected VS algorithms on the moving camera segmentation subtask.

evaluation is per-frame. This shows that the field of image segmentation is much more advanced than the field of video segmentation and performance on video segmentation can potentially be improved by transferring ideas from image segmentation. The proposed simple baseline method goes in this direction and immediately achieves the best performance on a wide working regime.

A further analysis regarding motion was performed by focusing on the camera motion. For the graphs in Figure 7 and the corresponding statistics in Table 2, we ignored all video sequences where the camera was not undergoing a considerable motion with respect to the depicted static 3D scene (video sequences with jitter were also not included). All algorithms positively maintained the same performance as on the general benchmark video set (cf. Figure 2 and Table 1), clearly indicating that a moving camera is not an issue for state-of-the-art video segmentation algorithms.

7. Conclusion and future work

In this work, we have addressed two fundamental limitations in the field video segmentation: the lack of a common dataset with sufficient annotation and the lack of an evaluation metric that is general enough to be employed on a large set of video segmentation subtasks. We showed that the dataset allows for an analysis of the current state-of-the-art in video segmentation, as we could address many working regimes - from over-segmentation to motion segmentation - with the same dataset and metric. This has led to interesting findings, for instance, that a quite simple baseline method can outperform all state-of-the-art methods in a certain working regime. This encourages progress on new aspects of the video segmentation problem. We have observed that the performance of the best performing method on all regimes is quite low for this complex dataset. This sets an important challenge in video segmentation and will foster progress in the field.

The proposed dataset is open to grow continuously and we welcome other researchers to contribute sequences and annotation to complement the dataset. Especially other subtasks could be included in the future by providing extra annotation.

Acknowledgements

We acknowledge partial funding by the ERC Starting Grant VideoLearn.

References

[1] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898–916, 2011.

[2] W. Brendel and S. Todorovic. Video object segmentation by tracking regions. In *ICCV*, 2009.

[3] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *PRL*, 30:88–97, 2009.

[4] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[5] A. Y. C. Chen and J. J. Corso. Propagating multi-class pixel labels throughout video frames. In *Western NY Image Worskshop*, 2010.

[6] H.-T. Cheng and N. Ahuja. Exploiting nonlocal spatiotemporal structure for video segmentation. In *CVPR*, 2012.

[7] J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille. Efficient multilevel brain tumor segmentation with integrated bayesian model classification. *TMI*, 27(5):629–640, 2008.

[8] D. DeMenthon. Spatio-temporal segmentation of video by hierarchical mean shift analysis. *SMVP*, 2002.

[9] K. Fragkiadaki and J. Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012.

[10] F. Galasso, R. Cipolla, and B. Schiele. Video segmentation with superpixels. In *ACCV*, 2012.

[11] F. Galasso, M. Iwasaki, K. Nobori, and R. Cipolla. Spatio-temporal clustering of probabilistic region trajectories. In *ICCV*, 2011.

[12] H. Greenspan, J. Goldberger, and A. Mayer. A probabilistic framework for spatio-temporal video representation. In *ECCV*, 2002.

[13] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010.

[14] A. Kannan, N. Jojic, and B. J. Frey. Generative model for layers of appearance and deformation. In *AISTATS*, 2005.

[15] M. P. Kumar, P. Torr, and A. Zisserman. Learning layered motion segmentations of video. (76):301–319, 2008.

[16] A. Levinshtein, C. Sminchisescu, and S. Dickinson. Spatiotemporal closure. In *ACCV*, 2010.

[17] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011.

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

[19] P. Ochs and T. Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011.

[20] S. Paris. Edge-preserving smoothing and mean-shift segmentation of video streams. In *ECCV*, 2008.

[21] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.

[22] A. Stein, D. Hoiem, and M. Hebert. Learning to find object boundaries using motion cues. In *ICCV*, 2007.

[23] P. Sundberg, T.Brox, M. Maire, P. Arbelaez, and J. Malik. Occlusion boundary detection and figure/ground assignment from optical flow. In *CVPR*, 2011.

[24] R. Tron and R. Vidal. A benchmark for the comparison of 3-D motion segmentation algoriths. In *CVPR*, 2007.

[25] D. Tsai, M. Flagg, and J. M. Rehg. Motion coherent tracking with multi-label mrf optimization. In *BMVC*, 2010.

[26] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *PAMI*, 2007.

[27] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Multiple hypothesis video segmentation from superpixel flows. In *ECCV*, 2010.

[28] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012.

[29] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *ECCV*, 2012.

[30] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *DAGM*, 2007.