

A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation

F. Perazzi^{1,2} J. Pont-Tuset¹ B. McWilliams² L. Van Gool¹ M. Gross^{1,2} A. Sorkine-Hornung²
¹ETH Zurich ²Disney Research

Abstract

Over the years, datasets and benchmarks have proven their fundamental importance in computer vision research, enabling targeted progress and objective comparisons in many fields. At the same time, legacy datasets may impend the evolution of a field due to saturated algorithm performance and the lack of contemporary, high quality data. In this work we present a new benchmark dataset and evaluation methodology for the area of video object segmentation. The dataset, named DAVIS (Densely Annotated VIdeo Segmentation), consists of fifty high quality, Full HD video sequences, spanning multiple occurrences of common video object segmentation challenges such as occlusions, motion-blur and appearance changes. Each video is accompanied by densely annotated, pixel-accurate and per-frame ground truth segmentation. In addition, we provide a comprehensive analysis of several state-of-the-art segmentation approaches using three complementary metrics that measure the spatial extent of the segmentation, the accuracy of the silhouette contours and the temporal coherence. The results uncover strengths and weaknesses of current approaches, opening up promising directions for future works.

1. Introduction

Video object segmentation is a binary labeling problem aiming to separate foreground object(s) from the background region of a video. A pixel-accurate, spatio-temporal bipartition of the video is instrumental to several applications including, among others, action recognition, object tracking, video summarization, and rotoscoping for video editing. Despite remarkable progress in recent years, video object segmentation still remains a challenging problem and most existing approaches still exhibit too severe limitations in terms of quality and efficiency to be applicable in practical applications, *e.g.* for processing large datasets, or video post-production and editing in the visual effects industry.

What is most striking is the performance gap among state-of-the-art video object segmentation algorithms and closely related methods focusing on image segmentation

Figure 1: Sample sequences from our dataset, with ground truth segmentation masks overlayed. Please refer to the supplemental material for the complete dataset.

and object recognition, which have experienced remarkable progress in the recent years. A key factor bootstrapping this progress has been the availability of large scale datasets and benchmarks [12, 26, 29, 42]. This is in stark contrast to video object segmentation. While several datasets exist for various different video segmentation tasks [1, 4, 5, 15, 20, 21, 25, 38, 41, 44, 46, 47], none of them targets the specific task of video *object* segmentation.

To date, the most widely adopted dataset is that of [47], which, however, was originally proposed for joint segmentation and tracking and only contains six low-resolution video sequences, which are not representative anymore for the image quality and resolution encountered in today's video processing applications. As a consequence, evaluations performed on such datasets are likely to be overfitted, without reliable indicators regarding the differences between individual video segmentation approaches, and the real performance on unseen, more contemporary data becomes difficult to determine [6]. Despite the effort of some authors to augment their evaluation with additional datasets, a standardized and widely adopted evaluation methodology for video object segmentation does not yet exist.

To this end, we introduce a new dataset specifically designed for the task of video object segmentation. The

dataset, which will be made publicly available, contains fifty densely and professionally annotated high-resolution Full HD video sequences, with pixel-accurate ground-truth data provided for every video frame. The sequences have been carefully captured to cover multiple instances of major challenges typically faced in video object segmentation. The dataset is accompanied with a comprehensive evaluation of several state-of-the-art approaches [5, 7, 13, 14, 18, 21, 24, 33, 35, 40, 43, 45]. To evaluate the performance we employ three complementary metrics measuring the spatial accuracy of the segmentation, the quality of the silhouette and its temporal coherence. Furthermore, we annotated each video with specific attributes such as *occlusions*, *fast-motion*, *non-linear deformation* and *motion-blur*. Correlated with the performance of the tested approaches, these attributes enable a deeper understanding of the results and point towards promising avenues for future research. The components described above represent a complete benchmark suite, providing researchers with the necessary tools to facilitate the evaluation of their methods and advance the field of video object segmentation.

2. Related Works

In this section we provide an overview of datasets designed for different video segmentation tasks, followed by a survey of techniques targeting video object segmentation.

2.1. Datasets

There exist several datasets for video segmentation, but none of them has been specifically designed for video *object* segmentation, the task of pixel-accurate separation of foreground object(s) from the background regions.

The *Freiburg-Berkeley Motion Segmentation* dataset [5] *MoSeg* is a popular dataset for motion segmentation, *i.e.* clustering regions with similar motion. Despite being recently adopted by works focusing on video object segmentation [35, 45], the dataset does not fulfill several important requirements. Most of the videos have low spatial resolution, segmentation is only provided on a sparse subset of the frames, and the content is not sufficiently diverse to provide a balanced distribution of challenging situations such as fast motion and occlusions.

The *Berkeley Video Segmentation Dataset* (BVSD) [44] comprises a total 100, higher resolution sequences. It was originally meant to evaluate occlusions boundary detection and later extended to over- and motion-segmentation tasks (VSB100 [19]). However, several sequences do not contain a clear object. Furthermore, the ground-truth, available only for a subset of the frames, is fragmented, with most of the objects being covered by multiple manually annotated, disjoint segments, and therefore this dataset is not well suited for evaluating video object segmentation.

SegTrack [47] is a small dataset composed of 6 densely annotated videos of humans and animals. It is designed to be challenging with respect to background-foreground color similarity, fast motion and complex shape deformation. Although it has been extensively used by several approaches, its content does not sufficiently span the variety of challenges encountered in realistic video object segmentation applications. Furthermore, the image quality is not anymore representative of modern consumer devices, and due to the limited number of available video sequences, progress on this dataset plateaued. In [25] this dataset was extended with 8 additional sequences. While this is certainly an improvement over the predecessor, it still suffers of the same limitations. We refer the reader to the supplemental material for a comprehensive summary of the properties of the aforementioned datasets, including ours.

Other datasets exist, but they are mostly provided to support specific findings and thus are either limited in terms of total number of frames, [8, 21, 25, 47], or do not exhibit a sufficient variety in terms of content [1, 4, 5, 15, 17, 20, 41, 46]. Others cover a broader range of content but do not provide enough ground-truth data for an accurate evaluation of the segmentation [21, 38]. Video datasets designed to benchmark tracking algorithms typically focus on surveillance scenarios with static cameras [9, 16, 32], and usually contain multiple instances of similar objects [50] (e.g. a crowd of people), and annotation is typically provided only in the form of axis-aligned bounding boxes, instead of pixel-accurate segmentation masks necessary to accurately evaluate video object segmentation. Importantly, none of the aforementioned methods includes contemporary high resolution videos, which is an absolute necessity to realistically evaluate the actual practical utility of such algorithms.

2.2. Algorithms

We categorize the body of literature related to video object segmentation based on the level of supervision required.

Unsupervised approaches have historically targeted over-segmentation [21, 51] or motion segmentation [5, 18] and only recently automatic methods for foreground-background separation have been proposed [13, 25, 33, 43, 45, 52]. These methods extend the concept of salient object detection [34] to videos. They do not require any manual annotation and do not assume any prior information on the object to be segmented. Typically they are based on the assumption that object motion is dissimilar from the surroundings. Some of these methods generate several ranked segmentation hypotheses [24]. While they are well suited for parsing large scale databases, they are bound to their underlying assumption and fail in cases it does not hold.

Semi-supervised video object segmentation methods propagate a sparse manual labeling, generally given in the form of one or more annotated frames, to the entire video

ID	Description
BC	<i>Background Clutter.</i> The back- and foreground regions around the object boundaries have similar colors (χ^2 over histograms).
DEF	<i>Deformation.</i> Object undergoes complex, non-rigid deformations.
MB	<i>Motion Blur.</i> Object has fuzzy boundaries due to fast motion.
FM	<i>Fast-Motion.</i> The average, per-frame object motion, computed as centroids Euclidean distance, is larger than $d_{fm} = 20$ pixels.
LR	<i>Low Resolution.</i> The ratio between the average object bounding-box area and the image area is smaller than $r_{lr} = 0.1$.
OCC	<i>Occlusion.</i> Object becomes partially or fully occluded.
OV	<i>Out-of-view.</i> Object is partially clipped by the image boundaries.
SV	<i>Scale-Variation.</i> The area ratio among any pair of bounding-boxes enclosing the target object is smaller than $r_{sv} = 0.5$.
AC	<i>Appearance Change.</i> Noticeable appearance variation, due to illumination changes and relative camera-object rotation.
EA	<i>Edge Ambiguity.</i> Unreliable edge detection. The average ground-truth edge probability (using [11]) is smaller than $p_e = 0.5$.
CS	<i>Camera-Shake.</i> Footage displays non-negligible vibrations.
HO	<i>Heterogeneous Object.</i> Object regions have distinct colors.
IO	<i>Interacting Objects.</i> The target object is an ensemble of multiple, spatially-connected objects (e.g. mother with stroller).
DB	<i>Dynamic Background.</i> Background regions move or deform.
SC	<i>Shape Complexity.</i> The object has complex boundaries such as thin parts and holes.

Table 1: List of video attributes and corresponding description. We extend the annotations of [50] (*top*) with a complementary set of attributes relevant to video *object* segmentation (*bottom*). We refer the reader to the supplementary material for the list of attributes for each in video in the dataset, and corresponding visual examples.

sequence. While being different from each other, they often solve an optimization problem with an energy defined over a graph structure [1, 40, 48]. To model long-range spatio-temporal connections some approaches use fully connected graphs [35], higher-order potentials [22]. The recent work of Märki *et al.* [31] efficiently approximates non-local connections minimizing the graph energy in bilateral space.

Supervised approaches assume manual annotation to be repeatedly added during the segmentation process, with a human correcting the algorithm results in an iterative fashion [2, 14, 49, 53]. These methods generally operate *online*, forward processing frames to avoid overriding of previous manual corrections. They guarantee high segmentation quality at the price of time-consuming human supervision, hence they are suited only for specific scenarios such as video post-production.

We evaluate a large set of the state-of-the-art approaches on our proposed dataset, providing new insights and several pointers to areas for future research.

3. Dataset Description

In this section we describe our new dataset DAVIS (Densely Annotated Video Segmentation) specifically designed for the task of video object segmentation. Exam-

ple frames of some of the sequences are shown in Figure 1. Based on experiences with existing datasets we first identify four key aspects we adhere to, in order create a balanced and comprehensive dataset.

Data Amount and Quality. A sufficiently large amount of data is necessary to ensure content diversity and to provide a uniformly distributed set of challenges. Furthermore, having enough data is crucial to avoid over-fitting and to delay performance saturation, hence guaranteeing a longer lifespan of the dataset [6]. The quality of the data also plays a crucial role, as it should be representative of the current state of technology. To this end, DAVIS comprises a total of 50 sequences, 3455 annotated frames, all captured at 24fps and Full HD 1080p spatial resolution. Due to the computational complexity being a major bottleneck in video processing, the sequences have a short temporal extent (about 2-4 seconds), but include all major challenges typically found in longer video sequences, see Table 1.

Experimental Validation. For each video frame, we provide pixel-accurate, manually created segmentation in the form of a binary mask. While we subdivide DAVIS into training- and a test-set to provide guidelines for future works, in our evaluation, we do not make use of the partition, and instead consider the dataset as a whole, since most of the evaluated approaches are not trained and a grid-search estimation of the optimal parameters would be infeasible due to the involved computational complexity.

Object Presence. Intuitively each sequence should contain at least one target foreground-object to be separated from the background regions. The clips in DAVIS contain either one single object or two spatially connected objects. We choose not to have multiple distinct objects with significant motion in order to be able to fairly compare segmentation approaches operating on individual objects against those that jointly segment multiple objects. Moreover, having a single object per sequence disambiguates the detection performed by methods which are fully automatic. A similar design choice made in [27] has been successfully steering research in salient object detection from its beginnings to the current state-of-the-art. To ensure sufficient content diversity, which is necessary to comprehensively assess the performance of different algorithms, the dataset spans four evenly distributed classes (*humans, animals, vehicles, objects*) and several actions.

Unconstrained Video Challenges. To enable a deeper analysis and understanding of the performance of an algorithm, it is fundamentally important to identify the key factors and circumstances which might have influenced it. Thus, inspired by [50] we define an extensive set of video attributes representing specific situations, such as fast-motion, occlusion and cluttered background, that typically pose challenges to video segmentation algorithms. Attributes are summarized in Table 1. They are not exclu-

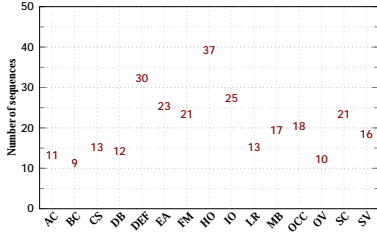


Figure 2: Left: *Attributes distribution over the dataset*. Each bin indicates the number of occurrences. Right: *Mutual dependencies among attributes*. The presence of a link indicates high probability of an attribute to appear in a sequence, if the one on the other end is also present.

sive, therefore a sequence can be annotated with multiple attributes. Their distribution over the dataset, *i.e.* number of occurrences, and their pairwise dependencies are shown in Figure 2. The annotations enable us to decouple the analysis of the performance into different groups with dominant characteristics (e.g. occlusion), yielding a better understanding of each methods’ strengths and weaknesses.

4. Experimental Validation

In order to judge the quality of a segmentation, the choice of a suitable metric is largely dependent on the end goal of the final application [10]. Intuitively, when video segmentation is used primarily a classifier within a larger processing pipeline, *e.g.* for parsing large scale datasets, it makes sense to seek the lowest amount of mislabeled pixels. On the other hand, in video editing applications the accuracy of the contours and their temporal stability is of highest importance, as these properties usually require the most painstaking and time-consuming manual input. In order to exhaustively cover the aforementioned aspects we evaluate the video segmentation results using three complementary error metrics. We describe the metrics in Section 4.1 and we empirically validate their complementary properties on the proposed dataset in Section 4.2.

4.1. Metrics Selection

In a supervised evaluation framework, given a ground-truth mask G on a particular frame and an output segmentation M , any evaluation measure ultimately has to answer the question how well M fits G . As justified in [37], for images one can use two complementary points of view, region-based and contour-based measures. As videos extends the dimensionality of still images to time, the temporal stability of the results must also be considered. Our evaluation is therefore based on the following measures.

Region Similarity J . To measure the region-based segmentation similarity, *i.e.* the number of mislabeled pixels, we employ the Jaccard index J defined as the *intersection-*

over-union of the estimated segmentation and the ground-truth mask. The Jaccard index has been widely adopted since its first appearance in PASCAL VOC2008 [12], as it provides intuitive, scale-invariant information on the number of mislabeled pixels. Given an output segmentation M and the corresponding ground-truth mask G it is defined as $J = \frac{|M \cap G|}{|M \cup G|}$.

Contour Accuracy F . From a contour-based perspective, one can interpret M as a set of closed contours $c(M)$ delimiting the spatial extent of the mask. Therefore, one can compute the contour-based precision and recall P_c and R_c between the contour points of $c(M)$ and $c(G)$, via a bipartite graph matching in order to be robust to small inaccuracies, as proposed in [28]. We consider the so called F-measure F as a good trade-off between the two, defined as $F = \frac{2P_c R_c}{P_c + R_c}$. For efficiency, in our experiments, we approximate the bipartite matching via morphology operators.

Temporal stability T . Intuitively, J measures how well the pixels of the two masks match, while F measures the accuracy of the contours. However, temporal stability of the results is a relevant aspect in video object segmentation—since the evolution of object shapes is an important cue for recognition and jittery, unstable boundaries are unacceptable in video editing applications. Therefore, we additionally introduce a temporal stability measure which penalizes such undesired effects.

The key challenge is to distinguish the *acceptable* motion of the objects from the undesired instability and jitter. To do so, we estimate the deformation needed to transform the mask at one frame to the next one. Intuitively, if the transformation is smooth and precise, the result can be considered stable.

Formally, we transform mask M_t of frame t into polygons representing its contours $P(M_t)$. We then describe each point $p_t^i \in P(M_t)$ using the Shape Context Descriptor (SCD) [3]. Next, we pose the matching as a Dynamic Time Warping (DTW) [39] problem, where we look for the matching between p_t^i and p_{t+1}^j that minimizes the SCD distances between the matched points while preserving the order in which the points are present in the shapes.

The resulting mean cost per matched point is used as the measure of temporal stability T . Intuitively, the matching will compensate motion and small deformations, but it will not compensate the oscillations and inaccuracies of the contours, which is what we want to measure. Occlusions and very strong deformations would be misinterpreted as contour instability, so we compute the measure on a subset of sequences without such effects.

4.2. Metrics Validation

To verify that the use of these measures produces meaningful results on our dataset, we compute the pairwise correlation between the region similarity J and the contour

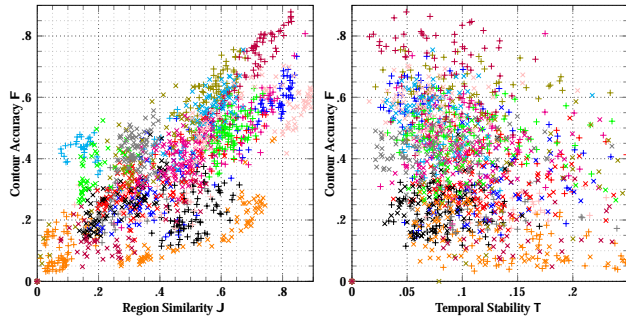


Figure 3: *Correlation between the proposed metrics.* Markers correspond to video frames. Colors encode membership to a specific video sequence. The contour accuracy measure F exhibits a slight linear dependency with respect to the region similarity J (left), while it appears uncorrelated to the temporal stability T (right).

accuracy F and between F and the temporal stability measure T . The degree of correlation is visualized in Figure 3. As can be expected, there is a tendency towards linear correlation between J and F (Figure 3, left), which can be explained by the observation that higher quality segmentations usually also result in more accurate contours. We note, however, that the level of independence is enough to justify the use of both measures. To get a qualitative idea of the differences between the two measures, Figure 4 shows two results of discrepant judgments between J and F . The temporal stability measure T and the contour accuracy F instead are nearly uncorrelated (Figure 3, right), which is also expected since temporal instability does not necessarily impact the per-frame performance.

Figure 4: *Discrepancy between metrics.* Ground truth in red and an example segmentation result in green. On the left, the result is penalized by J because in terms of number of pixels there is a significant amount of false negatives (head and foot), while with respect to the boundary measure F the missed percentage is lower. On the right the response of both measures is switched. The discrepancy in terms of pixels is low because the erroneous area is small, but the boundaries are highly inaccurate.

5. Evaluated Algorithms

We evaluate a total of twelve video segmentation algorithms, which we selected based on their demonstrated

state-of-the-art performance and source code availability, and two techniques commonly used for preprocessing. The source code was either publicly available or it was shared by the authors upon request.

Within the unsupervised category we evaluate the performance of NLC [13], FST [33], SAL [43], TRC [18], MSG [5] and CVOS [45]. The three latter approaches generate multiple segments per-frame, and therefore, as suggested in [5], we solve the bipartite graph matching that maximizes region similarity in terms of J to select the most similar to the target object. Among the semi-supervised approaches, SEA [40], JMP [14], TSP [7] and HVS [21] are initialized using the first-frame. HVS is meant for hierarchical over-segmentation, hence we search the hierarchy level and the corresponding segments that maximizes J of the first frame, keeping the annotation fixed throughout the entire video. FCP [35] uses a pair of annotated object proposals to initialize the classifiers. In our evaluation KEY [24] is deemed to be semi-supervised since we override their abjectness score and instead use the ground-truth to select the optimal hypotheses which is then refined solving a series of spatio-temporal graph-cuts.

The selected algorithms span the categories devised in Section 2 based on the level of supervision. However, interactive approaches with manual feedback could theoretically yield optimal results, and are not directly comparable with un- and semi-supervised approaches, since the number of user edits, *e.g.* strokes, should be also taken into account. Therefore we cast JMP [14] into a semi-supervised method that propagates masks to consecutive frames similar to SEA [40]. We reduce the number of categories in Table 2 and Table 3 accordingly.

Additionally we evaluate the performance of a salient object detector and the performance of an object proposal generator, as their output is a useful indicator with respect to the various video segmentation algorithms that are built upon them. We extract per-frame saliency from CIE-Lab images (SF-LAB, [34]) and from inter-frame motion (SF-MOT, [34]), while we use ground-truth to select the hypotheses of the object proposal generator (MCG, [36]) maximizing the per-frame Jaccard region similarity J .

6. Quantitative Evaluation

In this section we report the results of the fifteen evaluated approaches. We first provide different statistics evaluated for each of the three error measures (regions, contours, temporal), and then discuss evaluation results at the attribute level (*e.g.* performance with respect to appearance changes).

For each of the methods we kept the default parameters fixed throughout the entire dataset. Despite a considerable effort to speed-up the computation (parallelizing preprocessing steps such as motion estimation or extraction

		Preprocessing			Unsupervised							Semi-Supervised				
Measure		MCG	SF-LAB	SF-MOT	NLC	CVOS	TRC	MSG	KEY	SAL	FST	TSP	SEA	HVS	JMP	FCP
J	Mean M	0.724	0.173	0.532	0.641	0.514	0.501	0.543	0.569	0.426	0.575	0.358	0.556	0.596	0.607	0.631
	Recall O	0.912	0.075	0.672	0.731	0.581	0.560	0.636	0.671	0.386	0.652	0.388	0.606	0.698	0.693	0.778
	Decay D	0.026	-0.020	0.050	0.086	0.127	0.050	0.028	0.075	0.084	0.044	0.385	0.355	0.197	0.372	0.031
F	Mean M	0.654	0.218	0.452	0.593	0.490	0.478	0.525	0.503	0.383	0.536	0.346	0.533	0.576	0.586	0.546
	Recall O	0.781	0.052	0.440	0.658	0.578	0.519	0.613	0.534	0.264	0.579	0.329	0.559	0.712	0.656	0.604
	Decay D	0.046	-0.016	0.052	0.086	0.138	0.066	0.057	0.079	0.072	0.065	0.388	0.339	0.202	0.373	0.039
T	Mean M	0.652	0.758	0.637	0.356	0.243	0.327	0.250	0.190	0.600	0.276	0.329	0.137	0.296	0.131	0.285

Table 2: Overall results of region similarity (J), contour accuracy (F) and temporal (in-)stability (T) for each of the tested algorithms. For rows with an upward pointing arrow higher numbers are better (e.g., mean), and vice versa for rows with downward pointing arrows (e.g., decay, instability).

of boundary preserving regions) and to reduce the memory footprint (caching intermediate steps), several methods based on global optimization routines cannot be easily accelerated. Therefore, in order to be able to evaluate all methods with respect to each other, we were forced to down-sample the videos to 480p resolution. Due to the enormous processing power required, we performed experiments on different machines and partly on a cluster with thousands of nodes and heterogeneous CPU cores. Indicative runtimes are reported in the supplementary material.

The evaluation scripts, the input data, and the output results are made publicly available¹.

We exclude from the evaluation the first frame, which is used as ground-truth by semi-supervised methods, and the last frame which is not processed by some of the approaches. The overall results and considerations are reported in Section 6.1 and summarized in Table 2, while the attributes-based evaluation is discussed in Section 6.2 and summarized in Table 3.

6.1. Error Measure Statistics

For a given error measure C we consider three different statistics. Let $R = \{S_i\}$ be the dataset of video sequences S_i and let $\bar{C}(S_i)$ be the error measure average on S_i . The *mean* is the average dataset error defined as $M_C(R) = \frac{1}{|R|} \sum_{S_i \in R} \bar{C}(S_i)$. The *decay* quantifies the performance loss (or gain) over time. Let $Q_i = \{Q_i^1, \dots, Q_i^4\}$ be a partition of S_i in quartiles, we define the *decay* as $D_C(R) = \frac{1}{|R|} \sum_{Q_i \in R} \bar{C}(Q_i^1) - \bar{C}(Q_i^4)$. The *object recall* measures the fraction of sequences scoring higher than a threshold, defined as $O_C(R) = \frac{1}{|R|} \sum_{S_i \in R} \mathbb{1}_{\bar{C}(S_i) > \tau}$, with $\tau = 0.5$ in our experiments.

The region-based evaluation for all methods is summarized in Table 2. The best performing approach in terms of mean *intersection-over-union* is NLC [13] ($M_J = 0.641$), closely followed by FCP [35] ($M_J = 0.631$). However, the latter has better object recall O_J and less decay D_J . We report that, at the time of submission, our concurrent work

BVS [31] scored $M_J = 0.665$, therefore being the best performer in terms of region similarity, with the advantage of having the parameters tuned on this specific dataset.

With the exception of FCP [35], which solves a global optimization problem over a fully connected graph, the semi-supervised approaches TSP [7], SEA [40], HVS [21] and JMP [14] propagate the initial manual segmentation iteratively to consecutive frames and thus exhibit temporal performance decay as reflected in the results. To alleviate this problem, propagating using bigger steps and interpolating the results in-between can reduce the drift and improve the overall results [14]. TRC [18] and MSG [5] belong to a class of methods that uses motion segmentation as a prior, but the resulting over-segmentation of the object reflects negatively on the average performance. CVOS [45] uses occlusion boundaries, but still encounters similar issues. Differently from TRC and MSG, CVOS performs online segmentation. It scales better to longer sequences in terms of efficiency but experiences higher decay.

Aiming at detecting per-frame indicators of potential foreground object locations, KEY [24], SAL [43], and FST [33] try to determine prior information sparsely distributed over the video sequence. The prior is consolidated enforcing spatio-temporal coherence and stability by minimizing an energy function over a locally connected graph. While the local connectivity enables propagation of the segmentation similar to those of the semi-supervised approaches listed above, these methods suffer less decay as annotations are available at multiple different time frames.

Within the *preprocessing* category, the oracle MCG [36] is an informative upper-bound for methods seeking the best possible proposal per-frame. It has the highest region-based performance J and superior object recall M_J . The performance of MCG, also supported by the good performance of FCP and KEY that use concurrent object proposal generators, indicates that this could be a promising direction for more future research. As expected, in video sequences motion is a stronger low-level cue for object presence than color. Consequently salient motion detection SF-MOT [34] shows a significantly better performance than SF-LAB.

¹<https://github.com/fperazzi/davis>

Attr	Unsupervised							Semi-Supervised				
	NLC	CVOS	TRC	MSG	KEY	SAL	FST	TSP	SEA	HVS	JMP	FCP
AC	0.54 $+0.13$	0.42 $+0.12$	0.37 $+0.17$	0.48 $+0.08$	0.42 $+0.19$	0.33 $+0.12$	0.55 $+0.04$	0.17 $+0.23$	0.46 $+0.12$	0.42 $+0.23$	0.58 $+0.03$	0.51 $+0.16$
DB	0.53 $+0.15$	0.37 $+0.18$	0.39 $+0.15$	0.43 $+0.15$	0.52 $+0.07$	0.35 $+0.10$	0.53 $+0.06$	0.40 -0.06	0.58 -0.03	0.60 -0.01	0.60 $+0.01$	0.62 $+0.01$
FM	0.64 $+0.00$	0.37 $+0.24$	0.41 $+0.16$	0.46 $+0.14$	0.50 $+0.12$	0.35 $+0.13$	0.50 $+0.12$	0.18 $+0.31$	0.40 $+0.28$	0.42 $+0.31$	0.50 $+0.18$	0.55 $+0.13$
MB	0.61 $+0.04$	0.36 $+0.23$	0.32 $+0.27$	0.35 $+0.29$	0.51 $+0.08$	0.33 $+0.15$	0.48 $+0.14$	0.15 $+0.32$	0.39 $+0.24$	0.44 $+0.24$	0.51 $+0.15$	0.53 $+0.15$
OCC	0.70 -0.09	0.43 $+0.13$	0.44 $+0.10$	0.48 $+0.10$	0.52 $+0.08$	0.44 -0.02	0.53 $+0.07$	0.27 $+0.14$	0.47 $+0.13$	0.53 $+0.11$	0.47 $+0.21$	0.59 $+0.07$

Table 3: *Attribute-based aggregate performance.* For each method, the respective left column corresponds to the average region similarity J over all sequences with that specific attribute (e.g., AC), while the right column indicates the performance gain (or loss) for that method for the remaining sequences without that respective attribute. Only a subset of the most informative attributes from Table 1 are shown here. Please refer to the supplemental material for the complete evaluation.

The evaluation clearly shows that both the aggregate and individual performance of the approaches leave abundant room for future research. For instance, in [23] it is observed that a Jaccard index of $J = 0.7$ seems to be sufficiently accurate while $J = 0.6$ already represents a significant departures from the original object shape. The top techniques evaluated on DAVIS are still closer to the latter.

In terms of contour accuracy the best performing approaches are NLC and JMP. The former uses a large number of superpixels per-frame (≈ 2000) and a discriminative ensemble of features to represent them. In contrast, JMP exploits geodesic active contours to refine the object boundaries. The motion clusters of TRC and MSG, as well as the occlusion boundaries of CVOS generate sub-optimal results along the boundaries. The top ranked methods in terms of temporal stability are those that propagate segmentation on consecutive frames (JMP, SEA). As expected those that are used on a per-frame basis and cannot enforce continuity over time, such as MCG and SF-(*) generate considerably higher temporal instability. As a sanity check, we evaluate the temporal stability of the ground truth and we get $T = 0.093$, which is lower than any of the sequences.

6.2. Attributes-based Evaluation

As discussed in Section 3 and Table 1 we annotated the video sequences with attributes each representing a different challenging factor. These attributes allow us to identify groups of videos with a dominant feature *e.g.*, presence of occlusions, which is key to explaining the algorithms’ performance. However, since multiple attributes are assigned to each sequence, there might exist hidden dependencies among them which could potentially affect an objective analysis of the results. Therefore, we first conduct a statistical analysis to establish these relationship, and then detail the corresponding evaluation results.

Attributes Dependencies. We consider the presence or absence of each attribute in a video sequence to be represented as a binary random variable, the dependencies between which can be modelled by a pairwise Markov random field (MRF) defined on a graph G with vertex set

$V = \{1, \dots, 16\}$ and (unknown) edge set E . The absence of an edge between two attributes denotes that they are *independent* conditioned on the remaining attributes. Given a collection of $n = 50$ binary vectors denoting the presence of attributes in each video sequence, we estimate E via ℓ_1 penalized logistic regression. To ensure robustness in the estimated graph we employ *stability selection* [30]. Briefly, this amounts to performing the above procedure on $n/2$ -sized subsamples of the data multiple times and computing the proportion of times each edge is selected. Setting an appropriate threshold on this selection probability allows us to control the number of wrongly estimated edges according to Theorem 1 in [30]. For example, for a threshold value of 0.6 and choosing a value of k which on average selects neighbourhoods of size 4, the number of wrongly selected edges is at most 4 (out of $16^2 = 256$ possible edges). The estimated dependencies are visualized in Figure 2 (right). As expected there is a mutual dependency between attributes such as *fast-motion* (FM) and *motion-blur* (MB), or *interacting-object* (IO) and *shape-complexity* (SC). We refer the reader to the supplementary material for further details.

Results. In Table 3 we report the performance on subsets of the datasets characterized by a particular attribute. Due to space limitations we reduce the analysis in the paper to the most informative and recurrent attributes. Further details can be found in the supplementary material.

Appearance changes (AC) poses a challenge to several approaches, in particular for those methods strongly relying on color appearance similarity such as HVS and TCP. For example, TSP performance drops almost 50% as a consequence of the Gaussian process it uses to update the appearance model and therefore not being robust enough to strong appearance variations. Despite the dense connectivity of its conditional random field, FCP also experiences a considerable loss of performance. The reason resides in a sub-optimal automatic choice of the annotated proposals. Likely the proposals did have enough variety to span the entire object appearances causing the classifiers to overfit.

Dynamic background (DB) scenes, *e.g.* flowing water,

represent a major difficulty to the class of unsupervised methods, such as NLC and SAL, which adopt distinctive motion saliency as the underlying assumption to predict the object location. Interestingly the assumption of a completely closed motion boundary curve coinciding with the object contours can robustly accommodate background deformations (FST). Finally, MSG and TRC experience a considerable performance degradation as the motion clusters they rely on [5] are constructed from dissimilarities of point-trajectories, under the assumption that translational models are a good approximation for nearby points, which is not true on deforming image regions.

Fast motion (FM) is a problem for any of the algorithms exploiting motion information as the condition is a major challenge to reliable optical-flow computation. Note that there is a strong dependency between fast motion and motion-blur (MB) (Figure 2, *right*), yielding fuzzy object boundaries almost impossible to separate from the background region. Methods such as TRC and MSG use point-tracks for increased robustness towards fast motion, but are still susceptible with respect to motion-blur due to the sensitivity of the underlying variational approach used for densification of the results. NLC is the only method which has none or negligible loss of performance in both circumstances, possibly because the saliency computation is still reliable on a subset of the frames, and their random-walk matrix being non-locally connected is robust to fast motion.

Occlusions (OCC) being one of the well known challenges in video segmentation, only a small subset of the algorithms, which propagate sequentially manually annotated frames such as SEA and JMP, struggle with this type of situation. As expected, methods that exploit large range connectivity such as NLC, FCP and KEY are quite robust to these challenges.

7. Conclusion

To the best of our knowledge, this work represents the currently largest scale performance evaluation of video object segmentation algorithms. One of course has to consider that the evaluated approaches have been developed using different amounts and types of input data and ground-truth, or were partially even designed for different problems and only later adapted to the task of video object segmentation. However, the primary aim of our evaluation is not to determine a winner, but to provide researchers with high-quality, contemporary data, a solid standardized evaluation procedure, and valuable comparisons with the current state-of-the-art. We hope that the public availability of this dataset and the identified areas for potential future works will motivate even more interest in such an active and fundamentally important field for video processing.

As any dataset, also DAVIS will have a limited life-span. Therefore we welcome external contributions to extend it,

generalizing it to other segmentation tasks such as over-segmentation, or to other applications such as video alpha matting, semantic video segmentation, video retrieval, and action recognition.

Currently, running time efficiency and memory requirements are a major bottleneck for the usability of several video segmentation algorithms. In our experiments we observed that a substantial amount of time is spent pre-processing images to extract boundary preserving regions, object proposals and motion estimates. We encourage future research to carefully select those components bearing in mind they could compromise the practical utility of their work. Efficient algorithms will be able to take advantage of the Full HD videos and accurate segmentation masks made available with this dataset. Leveraging high resolution might not produce better results in terms of region-similarity, but it is essential to improve the segmentation of complex object contours and tiny object region.

Acknowledgements We thank the human and animal "actors" who contributed to the creation of DAVIS. In particular, we thank Lucia Colombo for her logistic support throughout the entire duration of the project. This work was partially funded by an SNF award (200021 143598).

References

- [1] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *CVPR*, 2010. 1, 2, 3
- [2] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: robust video object cutout using localized classifiers. *ACM Trans. Graph.*, 28(3), 2009. 3
- [3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *TPAMI*, 24(4), 2002. 4
- [4] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2), 2009. 1, 2
- [5] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 1, 2, 5, 6, 7
- [6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 1, 3
- [7] J. Chang, D. Wei, and J. W. F. III. A video representation using temporal superpixels. In *CVPR*, 2013. 2, 5, 6
- [8] A. Y. C. Chen and J. J. Corso. Propagating multi-class pixel labels throughout video frames. In *WNIIPW*, 2010. 2
- [9] R. Collins, X. Zhou, and S. K. Teh. An open source tracking testbed and evaluation web site. In *PETS 2005*, January 2005. 2
- [10] G. Csurka, D. Larlus, and F. Perronnin. What is a good evaluation measure for semantic segmentation? In *BMVC*, 2013. 4
- [11] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013. 3

- [12] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2), 2010. 1, 4
- [13] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014. 2, 5, 6
- [14] Q. Fan, F. Zhong, D. Lischinski, D. Cohen-Or, and B. Chen. Jumpcut: Non-successive mask transfer and interpolation for video cutout. *ACM Trans. Graph.*, 34(6), 2015. 2, 3, 5, 6
- [15] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 1, 2
- [16] R. B. Fisher. The pets04 surveillance ground-truth data sets. 2004. 2
- [17] K. Fragkiadaki and J. Shi. Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement. In *CVPR*, 2011. 2
- [18] K. Fragkiadaki, G. Zhang, and J. Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012. 2, 5, 6
- [19] F. Galasso, N. S. Nagaraja, T. J. Cardenas, T. Brox, and B. Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *ICCV*, 2013. 2
- [20] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *TPAMI*, 29(12), 2007. 1, 2
- [21] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010. 1, 2, 5, 6
- [22] S. D. Jain and K. Grauman. Supervoxel-consistent foreground propagation in video. In *ECCV*, 2014. 3
- [23] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *ECCV*, 2014. 7
- [24] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *Proc. ICCV*, 2011. 2, 5, 6
- [25] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013. 1, 2
- [26] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 1
- [27] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. *TPAMI*, 33(2), 2011. 3
- [28] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *TPAMI*, 26(5), 2004. 4
- [29] D. R. 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. 1
- [30] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. 7
- [31] N. Nicolas Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung. Bilateral space video segmentation. In *CVPR*, 2016. 3, 6
- [32] S. Oh, A. Hoogs, A. G. A. Perera, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011. 2
- [33] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013. 2, 5, 6
- [34] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 2, 5, 6
- [35] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung. Fully connected object proposals for video segmentation. In *ICCV*, 2015. 2, 3, 5, 6
- [36] J. Pont-Tuset, P. Arbeláez, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *TPAMI*, 2016. 5, 6
- [37] J. Pont-Tuset and F. Marques. Supervised evaluation of image segmentation and object proposal techniques. *TPAMI*, 2015. 4
- [38] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 1, 2
- [39] L. Rabiner and B.-H. Juang. Fundamentals of speech recognition. 1993. 4
- [40] S. A. Ramakanth and R. V. Babu. Seamseg: Video object segmentation using patch seams. In *CVPR*, 2014. 2, 3, 5, 6
- [41] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *CVPR Workshops*, 2009. 1, 2
- [42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 1
- [43] J. Shen, W. Wenguan, and F. Porikli. Saliency-Aware geodesic video object segmentation. In *CVPR*, 2015. 2, 5, 6
- [44] P. Sundberg, T. Brox, M. Maire, P. Arbelaez, and J. Malik. Occlusion boundary detection and figure/ground assignment from optical flow. In *CVPR*, 2011. 1, 2
- [45] B. Taylor, V. Karasev, and S. Soatto. Causal video object segmentation from persistence of occlusions. In *CVPR*, 2015. 2, 5, 6
- [46] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *CVPR*, 2007. 1, 2
- [47] D. Tsai, M. Flagg, and J. M. Rehg. Motion coherent tracking with multi-label MRF optimization. In *BMVC*, 2010. 1, 2
- [48] S. Vijayanarasimhan and K. Grauman. Active frame selection for label propagation in videos. In *ECCV*, 2012. 3
- [49] T. Wang, B. Han, and J. P. Collomosse. Touchcut: Fast image and video segmentation using single-touch interaction. *Computer Vision and Image Understanding*, 120, 2014. 3
- [50] Y. Wu, J. Lim, and M. Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 2, 3
- [51] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *CVPR*, 2012. 2
- [52] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013. 2
- [53] F. Zhong, X. Qin, Q. Peng, and X. Meng. Discontinuity-aware video object cutout. *ACM Trans. Graph.*, 31(6), 2012. 3