The 2017 DAVIS Challenge on Video Object Segmentation

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles,
Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool

Abstract—We present the *2017 DAVIS Challenge*, a public competition specifically designed for the task of video object segmentation. Following the footsteps of other successful initiatives, such as ILSVRC [1] and PASCAL VOC [2], which established the avenue of research in the fields of scene classification and semantic segmentation, the DAVIS Challenge comprises a dataset, an evaluation methodology, and a public competition with a dedicated workshop co-located with CVPR 2017. The DAVIS Challenge follows up on the recent publication of DAVIS (Densely-Annotated VIdeo Segmentation [3]), which has fostered the development of several novel state-of-the-art video object segmentation techniques. In this paper we describe the scope of the benchmark, highlight the main characteristics of the dataset and define the evaluation metrics of the competition.

Index Terms—Video Object Segmentation, DAVIS, Open Challenge, Video Processing

1 Introduction

Public benchmarks and challenges have been an important driving force in the computer vision field, with examples such as Imagenet [1] for scene classification and object detection, PASCAL [2] for semantic and object instance segmentation, or MS-COCO [4] for image captioning and object instance segmentation. From the perspective of the availability of annotated data, all these initiatives were a boon for machine learning researchers, enabling the development of new algorithms that had not been possible before. Their challenge and competition side motivated more researchers to participate and push towards the new different goals, by setting up a fair environment where test data are not publicly available.

The Densely-Annotated VIdeo Segmentation (DAVIS) initiative [3] provided a new dataset with 50 high-definition sequences with all their frames annotated with object masks at pixel-level accuracy, which has allowed the appearance of a new breed of video object segmentation algorithms [5], [6], [7], [8] that pushed the quality of the results significantly, almost getting to the point of saturation of the original dataset (around 80% performance by [5] and [6]). We will refer to this version of the dataset as DAVIS 2016.

To further push the performance in video object segmentation, we present *the 2017 DAVIS Challenge on Video Object Segmentation*, which consists of a new, larger, more challenging dataset (which we refer to as DAVIS 2017) and a public challenge competition and workshop. As the main new challenge, the new sequences have more than one annotated object in the scene, and we have re-annotated the original ones that have more than one visible object. The complexity of the videos has also increased with more distractors, smaller objects and fine structures, more occlusions and fast motion, etc. Overall, the new dataset consists of 150 sequences, totaling 10459 annotated frames and 376 objects. We will host a public competition challenge whose results will be presented in a workshop in the Computer Vision and Pattern Recognition (CVPR) conference 2017, Hawaii.

Figure 1 shows a set of example frames with the corresponding overlaid object annotations. The four first images come from newly-collected videos, while the latter from the DAVIS 2016 dataset re-annotated with multiple objects.

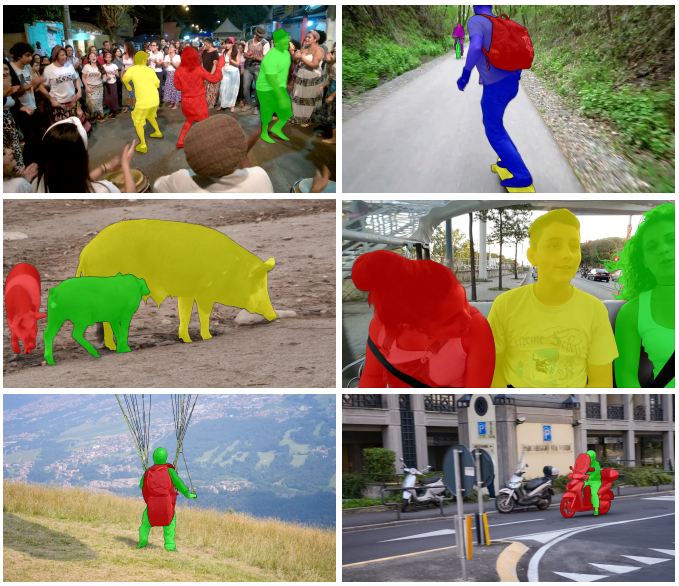


Fig. 1. Example annotations of the DAVIS 2017 dataset: The four first images come from new videos, the last two from videos originally in the DAVIS dataset re-annotated with multiple objects.

- *J. Pont-Tuset, S. Caelles, and L. Van Gool are with the Computer Vision Laboratory, ETH Zürich, Switzerland.*
- *F. Perazzi and A. Sorkine-Hornung are with Disney Research, Zürich, Switzerland.*
- *P. Arbeláez is with the Department of Biomedical Engineering, Universidad de los Andes, Colombia.*

Contacts and updated information can be found in the challenge website: http://davischallenge.org

	DAVIS 2016			DAVIS 2017				
	`train`	`val`	**Total**	`train`	`val`	`test-dev`	`test-challenge`	**Total**
Number of sequences	30	20	**50**	60	30	30	30	**150**
Number of frames	2079	1376	**3455**	4209	1999	2086	2180	**10474**
Mean number of frames per sequence	69.3	68.8	**69.1**	70.2	66.6	69.5	72.7	**69.8**
Number of objects	30	20	**50**	144	61	89	90	**384**
Mean number of objects per sequence	1	1	**1**	2.40	2.03	2.97	3.00	**2.56**

TABLE 1
Size of the DAVIS 2016 and 2017 dataset splits: number of sequences, frames, and annotated objects.

2 Dataset Facts and Figures

The main new challenge added to the DAVIS sequences in its edition of 2017 is the presence of multiple objects in the scene. As it is well known, the definition of an object is granular, as one can consider a person as including the trousers and shirt, or consider them as different objects. In DAVIS 2016 the segmented object was defined as the main *object* in the scene with a distinctive motion. In DAVIS 2017, we also segment the main moving objects in the scene, but we also divide them by semantics, even though they might have the same motion. Specifically, we generally segmented people and animals as a single instance, together with their clothes, (including helmet, cap, etc.), and separated any object that is carried and easily separated (such as bags, skis, skateboards, poles, etc.). As an example, Figure 2 shows different pairs of DAVIS 2016 segmentation (left) together with their DAVIS 2017 multiple-object segmentations.

Fig. 2. Example annotations of the DAVIS 2017 vs the single-object counterpart in DAVIS 2016: Semantics play a role even if the objects have the same motion.

As is a common practice in the computer vision challenges, we divide our dataset into different splits. First of all, we extend the `train` and `val` sets of the original DAVIS 2016, with annotations that will be made public for the whole sequence. We then define two other `test` sets (`test-dev` and `test-challenge`), for which only the masks on the first frames will be made public during the challenge. We will set up an evaluation server in Codalab where researchers will be able to submit their results, download an evaluation file, and publish their performance on the public leaderboard. For `test-dev` submissions will be unlimited and for a longer period of time, whereas `test-challenge`, which will determine the winners, will only be open for a short period of time and for a limited number of submissions.

Table 1 shows the number of sequences, frames, and objects on each of the dataset splits. Please note that `train` and `val` in DAVIS 2017 include the sequences of the respective sets in DAVIS 2016 with multiple objects annotated when applies. This is the reason why the mean number of objects per sequence is smaller in these two sets, despite all new sequences have around 3 objects per sequence in mean. The length of the sequences is kept similar to DAVIS 2016: around 70 frames.

In terms of resolution, the majority of new sequences are at 4k resolution (3840×2160 pixels), but there are also some 1440p, 1080p, and 720p images at its raw resolution. Despite this, the challenge will be on the downsampled 480p images, as it was the *de facto* standard for DAVIS 2016, and to facilitate their processing given the large amount of frames. We plan to increase the resolution used in future editions of the challenge.

3 Task Definition and Evaluation Metrics

The challenge will be focused on the so-called *semi-supervised* video object segmentation task, that is, the algorithm is given a video sequence and the mask of the objects in the first frame, and the output should be the masks of those objects in the rest of the frames. This excludes more supervised approaches that include a human in the loop (interactive segmentation) and unsupervised techniques (no initial mask is given). Please note that all objects in a frame have its unique identifier and so the expected output is a set of indexed masks by identifier.

Given a mask of a specific object given by an algorithm and the ground-truth mask of that same object in a specific frame, we use the region ($\mathcal{J}$) and boundary ($\mathcal{F}$) measures proposed in DAVIS 2016 [3]. Specifically, the former computes the number of pixels of the intersection between the two masks and divides it by the size of the union (also called Intersection over Union - IoU, or Jaccard index). The latter evaluates the accuracy in the boundaries, via a bipartite matching between the boundary pixels of both masks. The final boundary measure is the F measure between the precision and recall of the matching. Please refer to [3] for further description and discussion about these measures.

As of this edition, we discard the temporal instability ($\mathcal{T}$) given that its behavior is very affected by heavy occlusions.

In DAVIS 2016 we computed the measures on the subset of sequences with less occlusions but in DAVIS 2017 occlusions happen much more often, which would make the results less significant. Despite this, we encourage researchers to keep evaluating $\mathcal{T}$ and reporting it in the papers on the subset of selected sequences (available in the official code), since it is informative of the stability of the results.

As an overall measure of the performance of each algorithm we will compute the mean of the measures ($\mathcal{J}$ and $\mathcal{F}$) over all object instances. Formally, let S be a set of sequences, and O_S the set of annotated objects in these sequences. Given an object $o \in O_S$, $s(o) \in S$ is the sequence where the given object appears. Then, let F_s be the set of frames in sequence $s \in S$. Given a metric $\mathcal{M}$, the mean performance metric $m(\mathcal{M}, S)$ in the sequence set S is then defined as:

$$m(\mathcal{M}, S) = \frac{1}{|O_S|} \sum_{o \in O_S} \frac{1}{|F_{s(o)}|} \sum_{f \in F_{s(o)}} \mathcal{M}(m_o^f, g_o^f)$$

where m_o^f and g_o^f are the binary masks of the object and ground truth, respectively, of object o in frame f.

The overall performance metric that defines the ranking in a given set of the challenge is defined as:

$$M(S) = \frac{1}{2}\left[m(\mathcal{J}, S) + m(\mathcal{F}, S)\right]$$

as the average of the mean region and contour accuracies.

The performance of the metric in a given sequence $s \in S$ is defined as $m(\mathcal{M}, \{s\})$. Please note that we will report the metric per sequence as an informative measure, but the overall metric will not be the mean of the per-sequence values but per object as defined above, that is, in general $M(S) \neq \sum_{s \in S} M(\{s\})$.

4 Experiments

We will update this section and the web of the challenge (http://davischallenge.org) with the challenge results and an in-depth analysis of the results.

Acknowledgments

Research partially funded by the workshop sponsors: Google, Disney Research, NVIDIA, Prof. Luc Van Gool's Computer Vision Lab at ETHZ, and Prof. Fuxin Li's group at the Oregon State University. The authors gratefully acknowledge support by armasuisse, and thank NVIDIA Corporation for donating the GPUs used in this project.

References

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, 2015.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[3] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.

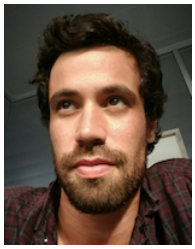
[5] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *CVPR*, 2017.

[6] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A.Sorkine-Hornung, "Learning video object segmentation from static images," in *CVPR*, 2017.

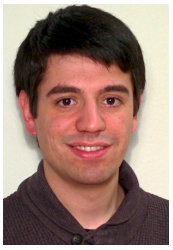
[7] N. Nicolas Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral space video segmentation," in *CVPR*, 2016.

[8] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung, "Fully connected object proposals for video segmentation," in *ICCV*, 2015.

Jordi Pont-Tuset is a post-doctoral researcher at ETHZ, Switzerland, in Prof. Luc Van Gool's Computer Vision Lab (since 2015). He received the degree in Mathematics in 2008, the degree in Electrical Engineering in 2008, the M.Sc. in Research on Information and Communication Technologies in 2010, and the Ph.D. with honors in 2014; from the Universitat Politècnica de Catalunya, BarcelonaTech (UPC). He worked at Disney Research, Zürich (2014).



Federico Perazzi is a post-doctoral researcher at Disney Research Zürich, Switzerland. He received the degree in Computer Science in 2008, the degree in Electrical Engineering in 2008, the M.Sc. in Computer Science and the Ph.D. from the Swiss Federal Institute of Technology (ETHZ), the M.Sc. in Entertainment Technology in 2010 from Carnegie Mellon University. He worked at Walt Disney Imagineering, where he developed the panoramic video stitching algorithm for the Disney Parks attraction "Soarin' Around the World". His research interests include computer vision, machine learning and computational photography.



Sergi Caelles is a Ph.D. student at ETHZ, Switzerland, in Prof. Luc Van Gool's Computer Vision Lab since 2016. He received the degree in Electrical Engineering and the M.Sc. in Telecommunications Engineering from the Universitat Politècnica de Catalunya, BarcelonaTech (UPC). He worked at Bell Laboratories, New Jersey (USA) in 2014. His research interest include computer vision with special focus on video object segmentation and deep learning.

Pablo Arbeláez received a Ph.D. with honors in Applied Mathematics from the Université Paris-Dauphine in 2005. He was a Research Scientist with the Computer Vision Group at U.C. Berkeley from 2007 to 2014. He currently holds a faculty position at Universidad de los Andes in Colombia. His research interests are in computer vision, where he has worked on a number of problems, including perceptual grouping, object recognition and the analysis of biomedical images.

Alexander Sorkine-Hornung is Senior Research Scientist at Disney Research Zürich, heading the Imaging and Video group. Before joining Disney, Alexander was a postdoctoral researcher at the Computer Graphics Laboratory at ETH Zürich. He obtained his Ph.D. in Computer Science at RWTH Aachen in 2008. Alexander's research interests lie in all areas related to digital image and video processing, at the interface of computer vision, graphics, and machine learning. In 2012 Alexander received the Eurographics Young Researcher Award. The research and technologies developed by his group have significant impact on Disney park attractions and movie productions, with film credits on movies such as Maleficent, Cinderella, and Big Hero 6.

Luc Van Gool got a degree in electromechanical engineering at the Katholieke Universiteit Leuven in 1981. Currently, he is professor at the Katholieke Universiteit Leuven, Belgium, and the ETHZ, Switzerland, Switzerland. He leads computer vision research at both places, where he also teaches computer vision. He has authored over 200 papers in this field. He has been a program committee member of several major computer vision conferences. His main interests include 3D reconstruction and modeling, object recognition, tracking, and gesture analysis. He received several Best Paper awards. He is a co-founder of 5 spin-off companies.