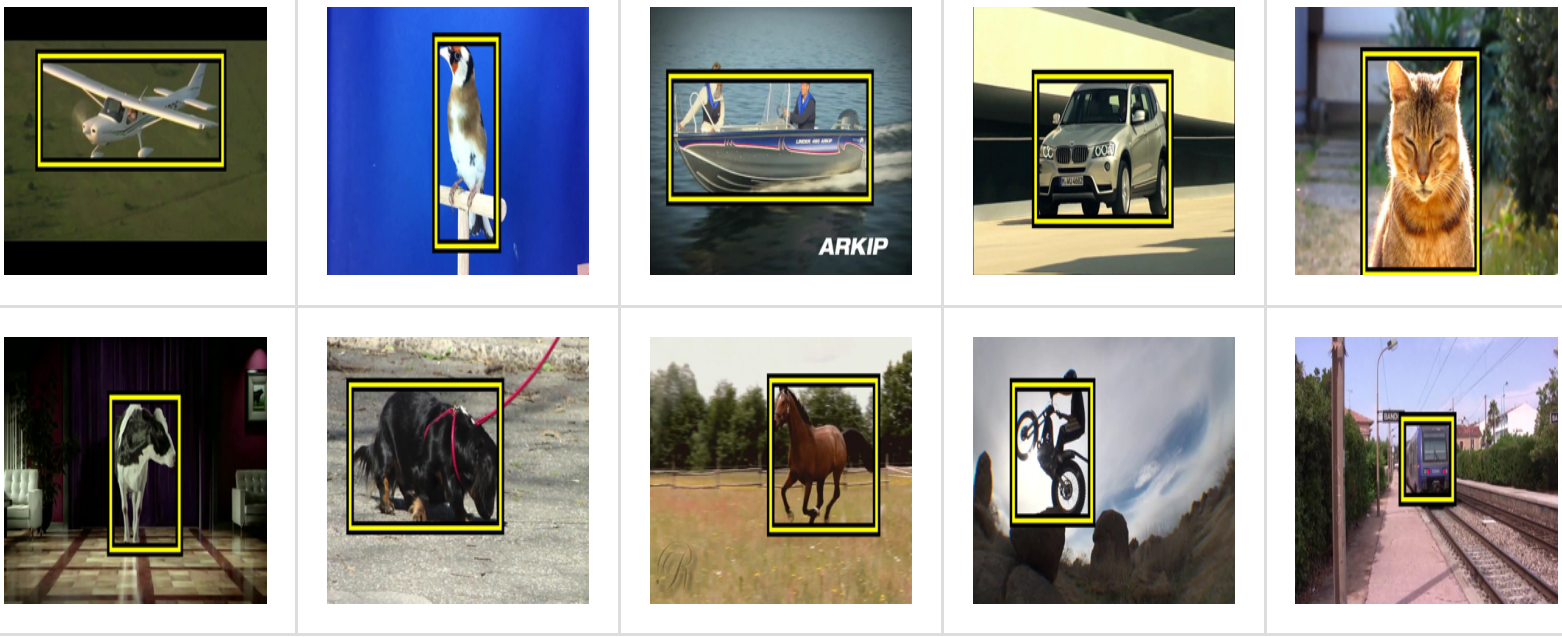
YouTube-Objects dataset v2.3

Alessandro Prest, Vicky Kalogeiton, Christian Leistner, Javier Civera, Cordelia Schmid, Vittorio Ferrari

University of Edinburgh (CALVIN), INRIA Grenoble (LEAR), ETH Zurich (CALVIN)

Overview



The YouTube-Objects dataset is composed of videos collected from YouTube by querying for the names of 10 object classes of the PASCAL VOC Challenge. It contains between 9 and 24 videos for each class. The duration of each video varies between 30 seconds and 3 minutes. The videos are weakly annotated, i.e. we ensure that each video contains at least one object of the corresponding class. If you use this dataset, please cite [1] and [3].

Dataset release v2.3

This release provides the annotations in PASCAL VOC 2007 format for the same 7,000 bounding box annotations from the YTO v2.2 [3].

You can download the dataset from here. If you use this release, please cite [3].

You can explore all annotated frames with the Dataset viewer.

Dataset release v2.2

In this release, we improved the quality of the images by fixing some decompression problems. As demonstrated in [3], we also have better shot boundaries and we have annotated more bounding boxes (6,975) than the ones contained in v1.0 (1,407).

The dataset contains a total of 720,000 frames. In order to eliminate possible confusion when decoding the videos and in the frame numbering, we release individual video frames after decompression.

classname	videos	shots	frames	classname	videos	shots	frames
aeroplane	13	482	79483	cow	11	70	41158
bird	16	175	34517	dog	24	217	86306
boat	17	191	119448	horse	15	151	70392
car	9	212	27607	motorbike	14	444	68421
cat	21	245	59822	train	15	324	132998

This release includes almost 7,000 bounding-box annotations [3]. For evaluation purposes we divided the annotated frames into training and test sets and we release them;in this manner, you are in possession of a perfect copy of the dataset as we used in our experiments [3]. In the training set, we annotated one instance per frame, while in the test set we annotated all instances of the desired object class.

classname	training instances	test images	test instances	classname	training instances	test images	test instances
aeroplane	415	170	180	cow	321	140	315
bird	359	155	162	dog	454	164	173
boat	357	144	234	horse	427	181	463
car	915	363	606	motorbike	360	165	213
cat	326	141	165	train	372	158	158

In addition to the videos and the bounding-box annotations, this release also includes several materials from our paper [3]:

- Original videos with the audio tracks.
- Optical flow, as produced by [4].
- Superpixels, as produced by [5].

Important Notice

These videos were downloaded from the internet, and may subject to copyright. We don’t own the copyright of the videos and only provide them for non-commercial research purposes.

Downloads v2.2

Filename	Description	Release Date	Size
README.txt	Description of contents	1 January 2015	6.0KB
Ranges.tar.gz	Videos and Shots of the dataset	1 January 2015	13.0KB
aeroplane.tar.gz		1 January 2015	1.8GB
bird.tar.gz		1 January 2015	3.0GB
boat.tar.gz		1 January 2015	11.0GB
car.tar.gz		1 January 2015	2.9GB
cat.tar.gz		1 January 2015	5.4GB
cow.tar.gz		1 January 2015	3.1GB
dog.tar.gz		1 January 2015	11.0GB
horse.tar.gz		1 January 2015	8.1GB
motorbike.tar.gz		1 January 2015	6.0GB
train.tar.gz		1 January 2015	11.0GB
GroundTruth.tar.gz	Ground truth annotations	7 April 2015	91.0KB
OpticalFlow.tar.gz	Optical flow by [4]	7 April 2015	4.3GB

SlicSuperpixels.tar.gz	Superpixels by [5]	17 April 2015	16.7GB
UsefulFiles.tar.gz	Useful files for the dataset	1 January 2015	27.0MB
YouTubeObjectsVideos.tar.gz	Videos (including audios)	1 January 2015	6.3GB

Dataset release v1.0

This release contains a total of 570'000 frames. As demonstrated in [1], the quality of the video frames play a crucial role in the performance of an object detector trained on them. We release individual video frames after decompression and after shot partitioning. In this manner, you are in possession of a perfect copy of the dataset as we used in our experiments [1].

classname	videos	shots	frames	classname	videos	shots	frames
aeroplane	13	1097	71327	cow	11	212	29642
bird	16	205	27532	dog	24	982	82432
boat	17	606	74501	horse	15	432	70247
car	9	208	14129	motorbike	14	511	40604
cat	21	220	42785	train	15	1034	117890

In addition to the videos, this release also includes several materials from our paper [1]

- Bounding-boxes annotations. For evaluation purposes we annotated the object location in a few hundred video frames for each class (see sec. 6.1 [1]).
- Point tracks and motion segments. As produced by [2].
- Tubes. Spatio-temporal bounding-boxes as described in section 3.2 [1]. We include all candidate tubes as well as the tube automatically selected by our method.

Downloads v1.0

Filename	Description	Release Date	Size
code.tar.gz	MATLAB source code to access the Youtube-Objects dataset.	17 June 2012	1MB
aeroplane.tar.gz		17 June 2012	2.0GB
bird.tar.gz		17 June 2012	3.0GB

boat.tar.gz		17 June 2012	7.6GB
car.tar.gz		17 June 2012	1.7GB
cat.tar.gz		17 June 2012	5.2GB
cow.tar.gz		17 June 2012	6.1GB
dog.tar.gz		17 June 2012	19.5GB
horse.tar.gz		17 June 2012	14.7GB
motorbike.tar.gz		17 June 2012	4.3GB
train.tar.gz		17 June 2012	21.1GB

References

1. Learning Object Class Detectors from Weakly Annotated Video
   Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, Vittorio Ferrari,
   In Computer Vision and Pattern Recognition (CVPR), 2012.
   
2. Object segmentation by long term analysis of point trajectories
   T. Brox, J. Malik,
   In European Conference on Computer Vision (ECCV), 2010.
   
3. Analysing domain shift factors between videos and images for object detection
   Vicky Kalogeiton, Vittorio Ferrari, Cordelia Schmid,
   In PAMI, 2016.
   
4. Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation
   Thomas Brox and Jitendra Malik
   In PAMI , 2011.
5. SLIC Superpixels Compared to State-of-the-art Superpixel Methods
   Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk
   In PAMI , 2012.

Acknowledgements

This work was partially funded by the QUAERO project supported by OSEO, French State

agency for innovation, the European integrated projects AXES and RoboEarth, DPI2009-07130, SNSF IZK0Z2-136096, CAIDGA IT 26/10, a Google Research Award and the ERC projects VisCul and ALLEGRO.

↑