ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding

Fabian Caba Heilbron[1,2], Victor Escorcia[1,2], Bernard Ghanem[2], Juan Carlos Niebles[1]
[1]Universidad del Norte, Colombia. [2]King Abdullah University of Science and Technology (KAUST), Saudi Arabia.

Figure 1: *ActivityNet* organizes a large number of diverse videos that contain human activities into a semantic taxonomy. **Top-row** shows the root-leaf path for the activity *Cleaning windows*. **Bottom-row** shows the root-leaf path for the activity *Brushing teeth*. Each box illustrates example videos that lie within the corresponding taxonomy node. Green intervals indicate the temporal extent of the activity. All figures are best viewed in color.

In spite of many dataset efforts for human action recognition, current computer vision algorithms are still severely limited in terms of the variability and complexity of the actions that they can recognize. This is in part due to the simplicity of current benchmarks, which mostly focus on simple actions and movements occurring on manually trimmed videos. In this paper we introduce *ActivityNet*, a new large-scale video benchmark for human activity understanding. Our benchmark aims at covering a wide range of complex human activities that are of interest to people in their daily living. In its current version, ActivityNet provides samples from 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 video hours. We illustrate three scenarios in which ActivityNet can be used to compare algorithms for human activity understanding: untrimmed video classification, trimmed activity classification and activity detection.

For example, note that the range of activities performed by one person in a day varies from making the bed after waking up to brushing teeth before going to sleep. Between these moments, he/she performs many activities relevant to his/her daily life. The American Time Use Survey reports that Americans spent an average 1.7 hours in household activities against only 18 minutes participating in sports, exercise or recreation per day [8]. In spite of this fact, most computer vision algorithms for human activity understanding are benchmarked on datasets that cover a limited number of activity types. In fact, existing databases tend to be specific and focus on certain types of activities such as: sports, cooking or simple actions. Typically, these datasets have a small number of categories (around 100), a small number of samples (short clips) per category (around 100), and limited category diversity.

In this paper, we address these dataset limitations by using a flexible framework that allows continuous acquisition, crowdsourced annotation, and segmentation of online videos, thus, culminating in a large-scale (large in the number of categories and number of samples per category), rich (diverse taxonomy), and easy-to-use (annotations, baseline classification models will be available online) activity dataset, known as *ActivityNet*.

We compare ActivityNet with several action datasets [1, 2, 3, 4, 5, 6, 7] in terms of: 1) variety in terms of the type of activities, and 2) number of activity classes and samples per class. To compare the variety on activity types, we manually annotate all the actions in each dataset with a parent top level category from the ActivityNet hierarchy. For example, the action *Push ups* from UCF101 is annotated under *Sports and exercising*. In Figure 2(Top), we plot a stacked histogram for the actions assigned to each top level category. It illustrates the lack of variety on activity types for all existing datasets. In contrast, ActivityNet strives for including activities in top level categories that are rarely considered in current benchmarks: *Household activities*, *Personal care*, *Education* and *Working activities*. To analyze the scale of ActivityNet compared to the existing action datasets, we plot in Figure 2(Bottom) the number of instances per class *vs* the number of activity/action classes. The current version of ActivityNet ranks second largest activity analysis dataset but it is the most varied in terms of activity types.

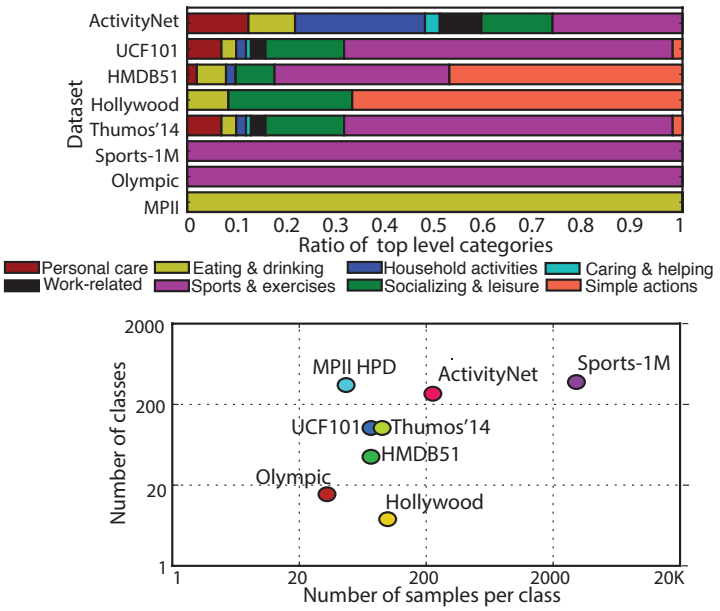

Figure 2: ActivityNet against existing datasets. Top compares the distribution of the activity classes in different datasets with the top levels of our hierarchy. Bottom compares the scale in terms of both number of samples per category and number of categories between different datasets.

Since a key goal of ActivityNet is to enable further development, research, and benchmarking in the field of human activity understanding, we are releasing our benchmark to the vision community. Annotations and a toolkit will be available at http://www.activity-net.org.

Acknowledgments We would like to thank the Stanford Vision Lab for their helpful comments and support. Research reported in this publication was supported by competitive research funding from King Abdullah University of Science and Technology (KAUST). JCN is supported by a Microsoft Research Faculty Fellowship.

[1] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[2] Amir Roshan Zamir Khurram Soomro and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. Technical report, University of Central Florida, 2012.

[3] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *ICCV*, 2011. doi: 10.1109/ICCV.2011.6126543.

[4] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[5] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[6] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[7] Thumos14. Thumos challenge 2014. http://crcv.ucf.edu/THUMOS14, 2013.

[8] U.S. Department of Labor. American time use survey. http://www.bls.gov/tus/, 2013.

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