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

Fabian Caba Heilbron[1,2], Victor Escorcia[1,2], Bernard Ghanem[2] and Juan Carlos Niebles[1]

[1]Universidad del Norte, Colombia
[2]King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Abstract

In this supplementary material, we complement our paper submission by providing additional analysis and results. First, we describe our taxonomy with more detail, providing the full hierarchy of ActivityNet. We also present screenshots of the crowdsourced annotation tools, and further statistics about data collection and annotation. Finally, we present more detailed results for the activity detection benchmark task.

1. Complete ActivityNet taxonomy

Figure 1 shows the full organizational taxonomy behind ActivityNet. In the main manuscript, Figure 3 only illustrates the sub-tree under *Household Activities* due to space restrictions. Here, we observe that ActivityNet organizes many more human activities under the sub-trees *Personal Care*, *Work-related Activities*, *Sports, Exercise and Recreation*, *Socializing, Relaxing and Leisure*, *Eating and Drinking* and *Caring and Helping*. In the current version, 203 categories are included in ActivityNet. This illustrates the high diversity in activity categories included in ActivityNet. The rich structure of our taxonomy is an important asset for algorithms that may exploit such information in activity analysis.

2. Collection and Annotation details

We illustrate the user interfaces for crowdsourced verification and temporal trimming of the human activities of interest.

Verifying videos Given a set of potential videos, we create tasks with batches of 15 videos which are sent to AMT workers. Figure 2 shows the instructions (2a) and the annotation interface (2b) provided to the turkers. In the instructions panel, we provide positive and negative examples that illustrate workers how to correctly perform the tasks. As shown, the task is designed to minimize worker effort.

Temporal trimming Figure 3 shows the user interface for annotating the temporal intervals that depict instances of the activity of interest. Users can easily navigate through the video and visualize the starting and ending frames of an activity instance. In order to promote consistency among annotations from different workers, we give a detailed description on how to annotate an activity instance.

3. Activity Detection: Precision-recall curves

Figure 4 shows example results of the activity detection task. We present the precision-recall curves for the four easiest (a-d) and hardest (e-h) classes. We find that activities with unclear temporal boundaries and those that contain complex human-object interactions are most difficult to detect. For example, the activity *Fixing mailbox* could include an interaction with a hammer or not. In contrast, activities related to sports tend to be confined to a temporal segment and usually involve similar scenarios and similar objects. In the case of *Pole vault* the movements of the athlete tend to be highly structured and a flexible pole is always present in the video.

4. Complexity analysis

In this section, we report the time required for computing visual features from the ActivityNet videos. As shown in Table 4, computing motion features requires about 6 years of processing in a single modern CPU core. This evidences the need for novel algorithms that provide efficient video description and feature extraction.

Feature type	Codebook	Extraction	Encoding	Total
MF	144	50400	480	51024
DF	NA	84	NA	84
SF	42	2050	103	2195
MF+DF+SF	186	52534	583	53303

Table 1: Time in hours spent on feature computation. It includes the time required to save the data on disk.

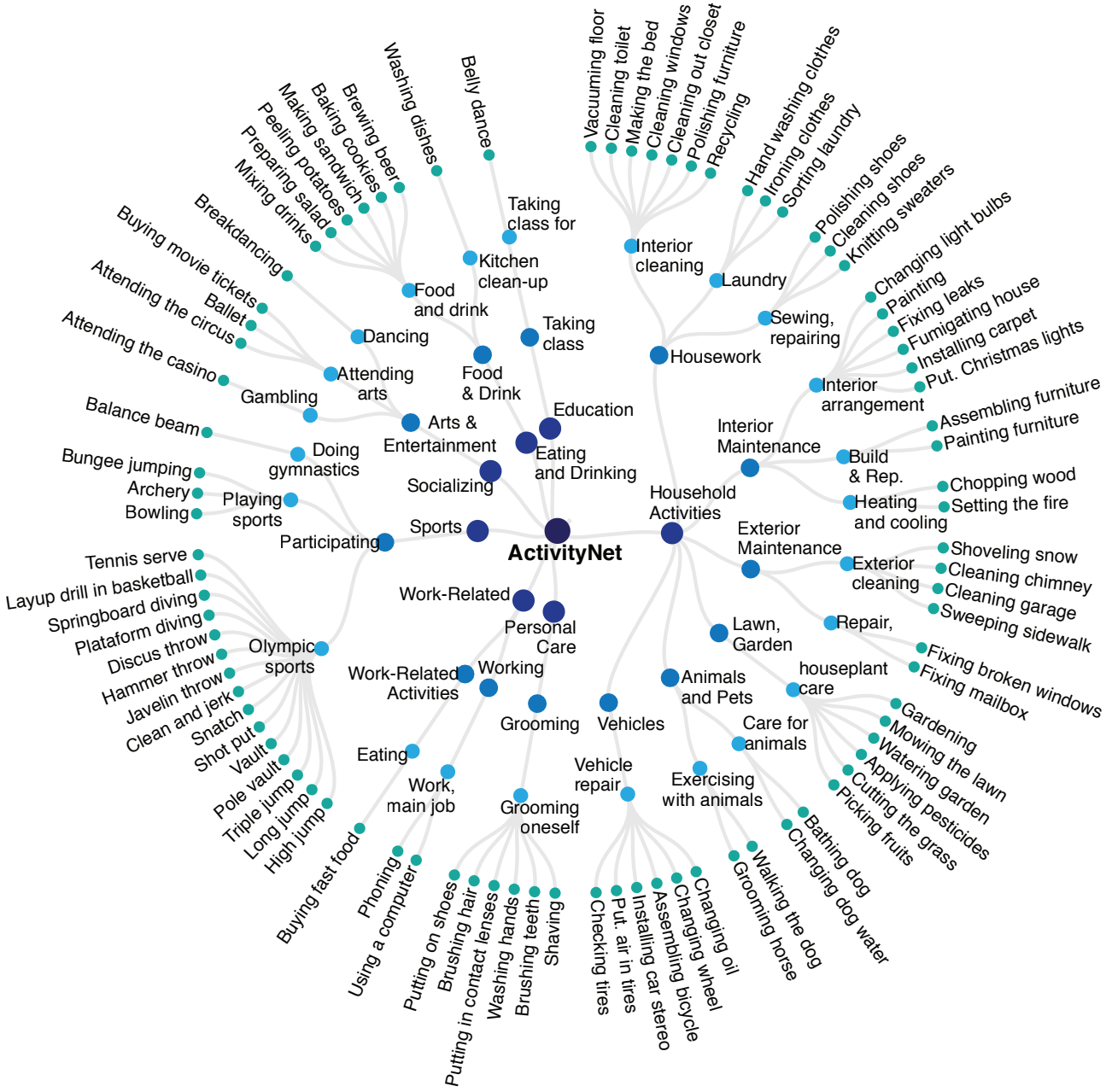
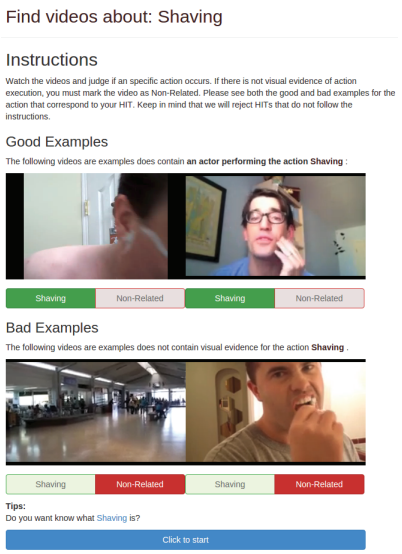
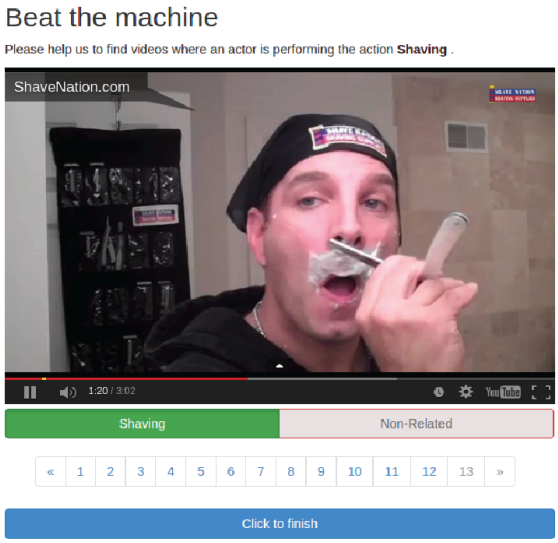


Figure 1: Complete organizational taxonomy behind ActivityNet.

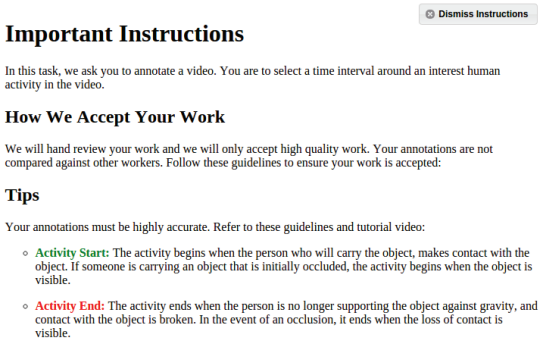


(a) Instructions for verifying activities

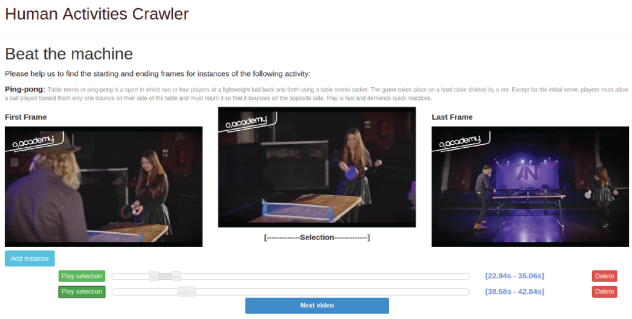


(b) User interface for annotating presence of activities

Figure 2: User interface for verifying human activities.

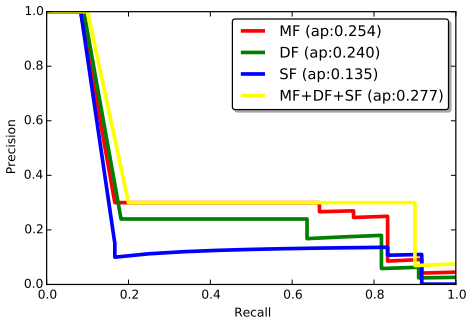


(a) Instructions for trimming activities

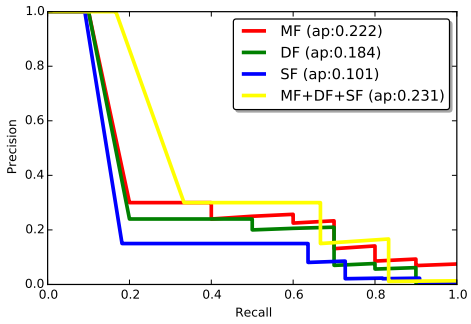


(b) User interface for trimming activity instances

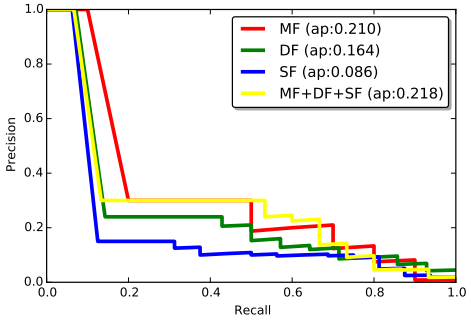
Figure 3: User interface for temporal annotation of human activity instances.



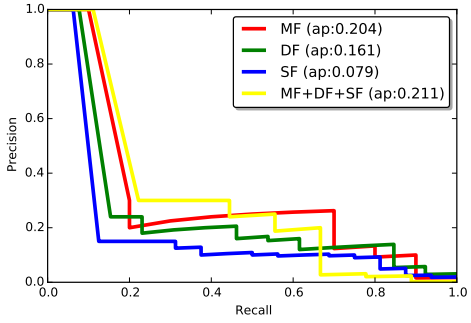
(a) Pole vault



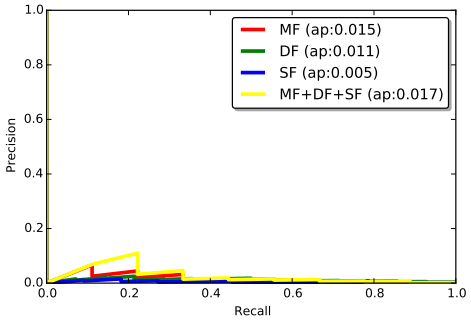
(b) Platform diving



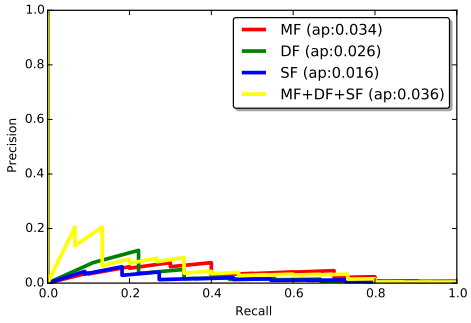
(c) Playing guitar



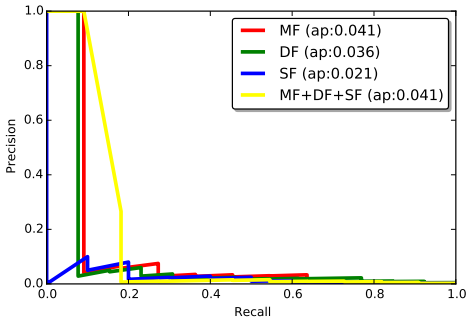
(d) Washing hands



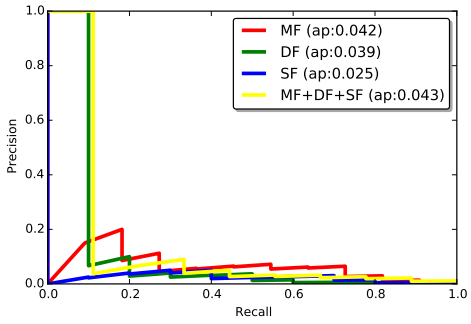
(e) Fixing mailbox



(f) Shoveling snow



(g) Organizing file cabinet



(h) Preparing shower for a child

Figure 4: Detection results.