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

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

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 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.*

1. Introduction

With the growth of online media, surveillance and mobile cameras, the amount and size of video databases are increasing at an incredible pace. For example, YouTube reported that over 300 hours of video are uploaded every minute to their servers [43]. Arguably, people are the most important and interesting subjects of such videos. The computer vision community has embraced this observation to validate the crucial role that human activity/action recognition plays in building smarter surveillance systems, semantically aware video indexes, and more natural human-computer interfaces. However, despite the explosion of video data, the ability to automatically recognize and understand human activities is still rather limited. This is primarily due to impeding challenges inherent to the task, namely the large variability in execution styles, complexity of the visual stimuli in terms of camera motion, background clutter and viewpoint changes, as well as, the level of detail and number of activities that can be recognized. An important limitation that hinders the performance of current techniques is the state of existing video datasets and benchmarks available to action/activity recognition researchers.

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 [37]. 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 *i.e*. 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*. One of the most important aspects of ActivityNet is that it is structured around a semantic ontology which organizes activities according to social interactions and where they usually take place. It provides a rich activity hierarchy with at least four levels of depth. For example, the activity *Filing nails* falls under the third tier category *Washing, dressing and grooming*, which belongs to the second tier *Grooming* and finally the major category *Personal care*. Figure 1 illustrates other examples of this organization. To the best of our knowledge, ActivityNet is the first database for human activity recognition organized under a rich semantic taxonomy.

We organize the paper as follows: we first review and summarize existing benchmarks for human activity understanding. Then, we present the details of our dataset collection and annotation framework and provide a summary of the properties of ActivityNet. We illustrate three

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.

benchmarking scenarios for evaluating the performance of state-of-the-art algorithms: untrimmed video classification, trimmed activity classification and activity detection.

2. Related Work

The challenges of building systems that understand and recognize complex activities in real environments and conditions, has prompted the construction of standardized datasets for algorithm training and evaluation. However, current benchmarks are rather limited in at least one of these aspects: number of categories, samples per category, temporal length of each sample, diversity of video capturing conditions or environments, and the diversity of category taxonomy. Furthermore, extending most of these datasets involves extremely costly manual labor.

We briefly review some of the most influential action datasets available. The Hollywood dataset [20] contains videos taken from Hollywood movies. Twelve action categories are performed by professional actors, which results in more natural scenes than earlier simple action datasets [33, 9]. Similarly, other datasets also relax the environment assumptions leading to challenging recognition tasks with difficult background and camera angles. For example, UCF Sports [30] and Olympic Sports [24] increase the action complexity by focusing on highly articulated sporting activities. However, the small number of categories keeps the scope of the activities narrow, and cannot be considered a representative sample of activities in the real-world. Another dimension of complexity is addressed by datasets that focus on composable [21] and concurrent [41] activities, but these are constrained with respect to the scene and environment assumptions.

Next in terms of sample size are the UCF101[17]-Thumos'14[35] and the HMDB51[19] datasets, compiled from YouTube videos and with more than 50 action categories. The resulting video samples are short and only convey simplistic short-term actions or events. These videos were collected through a manual and costly process, which is difficult to scale if the size of the dataset is to be extended. In terms of semantic organization, HMDB51 groups activities into 5 major types: general-facial, facial with object manipulation, general body movement, body movements with object interaction and body movements for human interaction. On the other hand, UCF101 groups categories into 5 types: human-object interaction, body motion only, playing musical instruments, sports. Unfortunately, these are simple taxonomies with only two levels of resolution, and do not provide a detailed organization of activities.

The MPII Human Pose Dataset [2] focuses on human pose estimation, and was recently applied to action recognition [29]. It provides short clips (41 frames or longer) that depict human actions. Unfortunately, the distribution of video samples per category is non-uniform and biased towards some action categories.

Currently, the largest video dataset available is the Sports-1M dataset [16], with about 500 sports-related categories, annotated by an automatic tagging algorithm. Despite its sheer size, this dataset is structured using a somewhat limited activity taxonomy, as it only focuses on sports actions. Furthermore, the automatic collection process introduces an undisclosed amount of label noise.

Also related to our work are the efforts to construct large-scale benchmarks for object recognition in static images. Image benchmarks such as ImageNet[5], SUN[42] and Tiny Images[36] have spawned significant advances for computer vision algorithms in the related tasks. An example is the Large Scale Visual Recognition Challenge (ILSVRC) [32], from which the *AlexNet* architecture [18] gains its popularity due to an outstanding performance in the challenge.

ActivityNet attempts to fill the gap in the following aspects: a large-scale dataset that covers activities that are most relevant to how humans spend their time in their daily living; a qualitative jump in terms of number and length

of each video (instead of short clips), diversity of activity taxonomy and number of classes; a human-in-the-loop annotation process that can provide higher label accuracy as compared to fully automatic annotation algorithms; and a framework for continuous dataset expansion at low cost.

3. Building ActivityNet

ActivityNet aims at providing a semantic organization of videos depicting human activities. In this section, we introduce the activity lexicon and hierarchy that serves as a backbone for ActivityNet. Another important goal is to provide a large set of diverse video samples for each activity of interest. In this section, we also describe our scalable data collection and video annotation scheme. Finally, we summarize some interesting properties of ActivityNet.

3.1. Defining the Activity lexicon

Our goal is to build ActivityNet upon a rich semantic taxonomy. In contrast to the object domain, it is difficult to define an explicit semantic organization of activities. Beyond the shallow hierarchies that organize current benchmarks, some attempts have been made at providing a structured organization of activities within the computer vision community. Aloimonos *et al*. [10, 26] propose a two-level organization of activities into 6 groups: ground, general object, general person, specific object, specific person, group; which connects to verbs in WordNet. Unfortunately, verbs are more difficult to use directly, because unlike objects in ImageNet [5], there is more ambiguity and polysemism between verbs and activities, than between objects and synsets. This may be partly explained by the fact that our spoken language for activities needs more complicated constructions compared to what is needed for objects.

Outside the vision community, there are efforts that organize general knowledge into structured repositories, such as Freebase[8], FrameNet[7], among others. Since none of them are specific to activities, their richness and depth are limited. On the other hand, there are also efforts more specific to activities. In the medical community, Ainsworth *et al*. [1] organizes a small number of physical human activities into a two level taxonomy.

Since we aim at a large scale benchmark with high activity diversity, we propose the use of the activity taxonomy built by the Department of Labor for conducting the American Time Use Survey [37]. The ATUS taxonomy organizes more than 2000 activities according to two key dimensions: a) social interactions and b) where the activity usually takes place. The ATUS coding lexicon contains a large variety of daily human activities organized under 18 top level categories such as Personal Care, Work-Related, Education and Household activities. In addition, there are two more levels of granularity under these top level categories. For example, the activity *Polishing shoes*, appears in the hierarchy

Figure 3. Visualization of the sub-tree of the top level category *Household activities*. Full taxonomy is available in the supplementary material.

as a leaf node under the third category, *Sewing, repairing and maintaining textiles*, which is part of the second tier category, *Housework*, which falls under the *Household* activities top level category.

For the first release of ActivityNet, we have manually selected a subset of 203 activity categories, out of the more than two thousand activity examples provided by the ATUS activity hierarchy. The activity classes belong to 7 different top level categories: *Personal Care*, *Eating and Drinking*, *Household*, *Caring and Helping*, *Working*, *Socializing and Leisure* and *Sports and Exercises*. Figure 3 illustrates the sub-tree for the top-level category *Household activities*. The rich taxonomy in ActivityNet, which has four levels of granularity, constitutes a semantic organization backbone that may be useful in algorithms that are able to exploit the hierarchy during model training.

3.2. Collecting and annotating human activities

Building benchmark datasets for visual recognition has been traditionally a difficult and time consuming task. The goal of ActivityNet is to provide a large-scale dataset of activities that can be expanded and annotated continously at a reasonably low cost. Traditional data collection practices that require many expert researcher hours are prohibitive. On the other hand, fully automatic methods introduce label noise that is difficult to erradicate.

We now describe the collection and annotation process for obtaining ActivityNet. Inspired by [5, 11, 38], we follow a semi-automatic crowdsourcing strategy to collect and annotate videos (Figure 2). We first search the web for potential videos depicting a particular human activity. Then, we

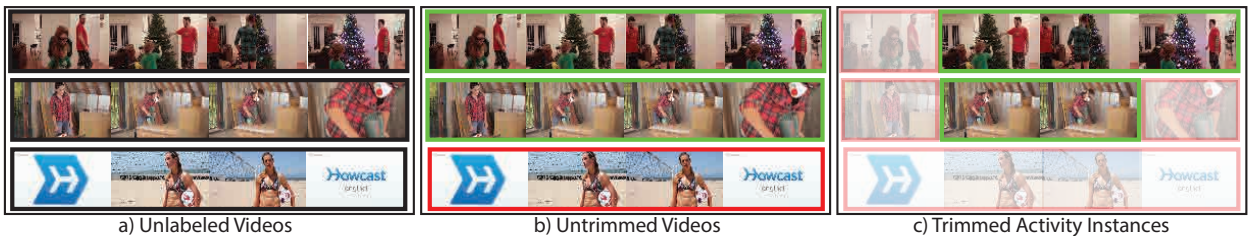

Figure 2. Video collection and annotation process. (a) We start with a large number of candidate videos, for which the labels are partially unknown. (b) AMT workers verify if an activity of interest is present in each video, so that we can discard false positive videos (in red). This results in a set of *untrimmed videos* that contain the activity (in green). (c) Finally, we obtain temporal boundaries for *activity instances* (in green) with the help of AMT workers.

rely on Amazon Mechanical Turk (AMT) workers to verify the presence of the activity in each video. Finally, multiple workers annotate each video with the temporal boundaries associated to the activity.

Search the Web: At this stage, we have a textual list of human activity classes and our goal is to search the web to retrieve videos related to each activity. Exploiting the large amount of video data on online repositories such as YouTube, we search videos using text based queries. These queries are expanded with WordNet [23] using hyponyms, hypernyms and synonyms in order to increase the number of retrieved videos and content variety.

Labeling Untrimmed Videos: We verify all videos retrieved and remove those not related with the activity at hand. We employ AMT workers (turkers) to review each video and determine if it contains an intended activity class. In order to keep the annotation quality high, we insert verifiable labeling questions and only employ multiple expert turkers. Due to the inaccuracy of text-based queries, we usually discard many videos that are not related with any of the intended activity classes. At the end of this process, we have a set of verified untrimmed videos that are associated to at least one ground truth activity label.

Annotating the Activity Instances: Most current activity classification systems require training videos to be trimmed to only contain the intended activity. Nevertheless, it is hard to find web videos containing only information with a specific activity. For example, when searching YouTube with the query "Preparing pasta", results include videos containing contextual information about the chef. In this direction, we aim to manually annotate the temporal boundaries where an activity is performed in a video. To tackle this manual process, we rely on AMT workers to temporally annotate all the activity instances present in a video. In order to ensure quality, the temporal extent of each activity instance is labelled by multiple expert turkers. Then, we cluster their annotations to obtain robust annotation agreements. This stage produces a curated set of activity instances, each of them associated to exactly one ground truth activity label. Moreover, it is important to note that within one untrimmed video, there may be more than one activity instance from more than one activity class.

3.3. ActivityNet at a Glance

We now look into some of the properties of the videos in ActivityNet. We first report statistics related to the video data. Second, we compare ActivityNet to several existing datasets for the benchmarking of human activities.

Video Properties All ActivityNet videos are obtained from online video sharing sites. We download the original videos at the best quality available. In order to limit the total storage requirement, we prioritize the search toward videos less than 20 minutes long. In practice, a large proportion of videos have a duration between 5 and 10 minutes. Around 50% of the videos are in HD resolution (1280×720), while the majority have a frame rate of 30 FPS.

Collection and Annotation Summary Figure 4-(top rows) shows the number of untrimmed videos and trimmed activity instances per class in the current version of ActivityNet. The distribution is close to uniform, which helps to avoid data unbalance when training classifiers. Also note that there is a factor of 1.41 trimmed instances per untrimmed video in average. Finally, our collection process will allow easy expansion of ActivityNet in terms of number of samples per category and number of categories.

Comparison with existing datasets We compare ActivityNet with several action datasets [17, 19, 20, 24, 35, 16, 31] 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 4(bottom-left), 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 4(bottom-right) 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.

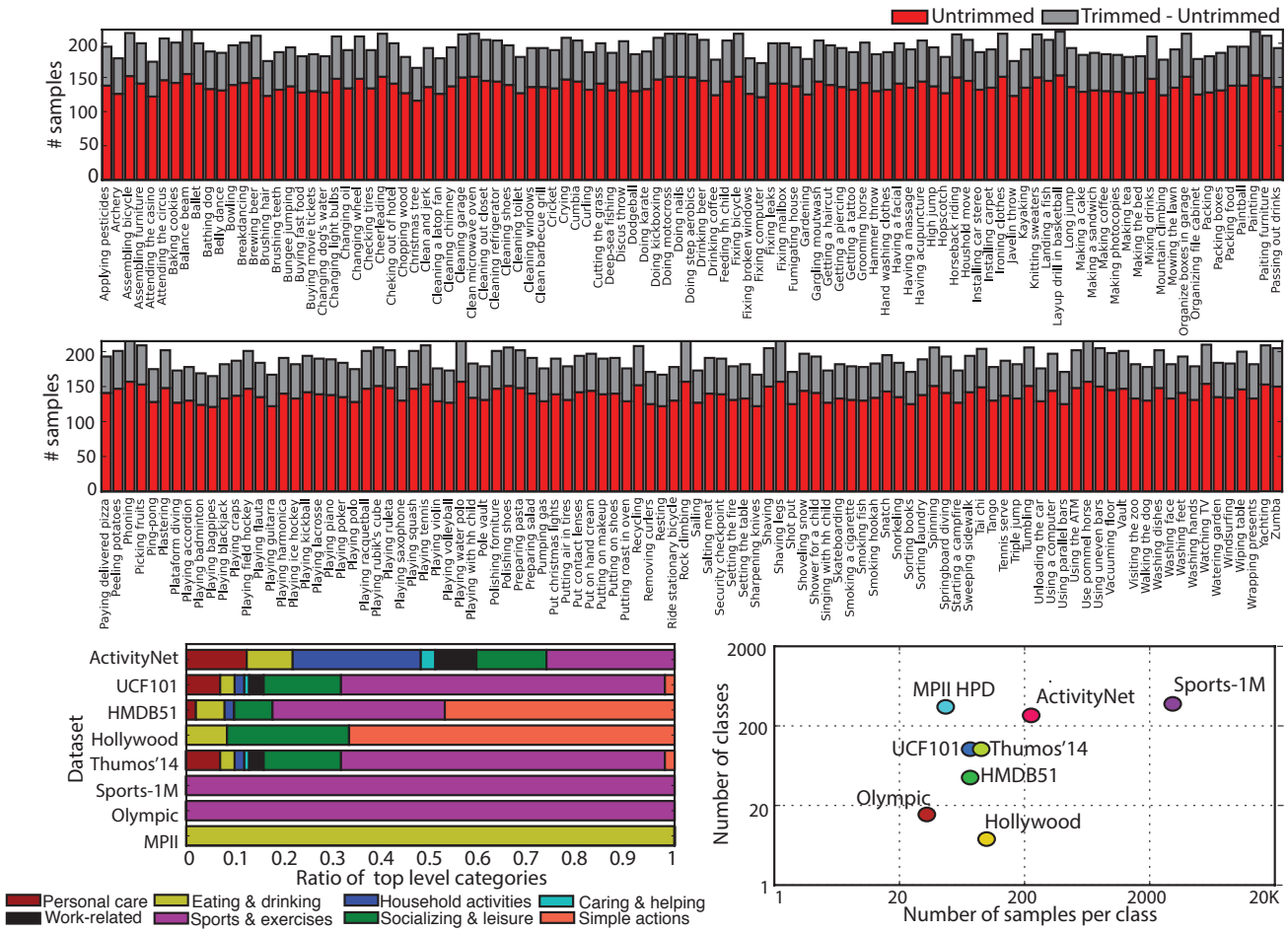

Figure 4. Summary of ActivityNet statistics. First two rows: Red bars indicate the number of untrimmed instances per activity category; and the top of the gray bars indicate the number of trimmed instances in each class. Bottom left compares the distribution of the activity classes in different datasets with the top levels of our hierarchy. Bottom right compares the scale in terms of both number of samples per category and number of categories between different datasets.

4. Experimental Results

This section presents a series of evaluations that showcase several benchmarking scenarios in which ActivityNet can be used. These evaluations also serve to illustrate the challenge that general activity understanding is to current computer vision algorithms.

The rest of the section is organized as follows. We first describe the video representations adopted in our evaluation scenarios. Then, we examine the performance of a state-of-the-art algorithm for action recognition in three different tasks: a) *Untrimmed video classification*, b) *Trimmed activity classification*, and c) *Activity detection*. For this study we choose the state-of-the-art action recognition pipeline from [25], which consists of improved trajectories, static and deep features encoded using fisher vectors, and a one-vs-all linear SVM as classifier. Lastly, we provide a cross-task analysis and discussions about the results obtained.

4.1. Video Representation

In order to capture visual patterns in each input video, we construct a video representation using a combination of several feature types: *motion features*, *static features* and *deep features*. This is motivated by the observation that combining multiple feature types can lead to significant improvements in action recognition [34, 25].

Motion Features (MF): These features aim to capture local motion patterns in a video. In practice, we first extract improved trajectories [39] to obtain a set of local descriptors *i.e*. HOG, HOF, and MBH. We encode these descriptors using the Fisher vector (FV) coding scheme [28], where each descriptor type is represented separately. In all our experiments, we first learn a GMM with 512 components and reduce the dimensionality of the final encoding to half using PCA. This is similar to the procedures in [39, 3, 4].

Static Features (SF): These features aim to encode contextual scene information. These context cues are usually helpful to discriminate human activities [12, 13]. In prac-

tice, we capture contextual scene information by extracting SIFT features every ten frames. These features are encoded using FV with a GMM of 1024 components, which is then reduced to a feature size of 48 dimensions using PCA. The final representation for each video aggregates all descriptors in a single FV.

Deep Features (DF): These features aim to encode information about the objects in the scene. In many activities involving object interactions, this is an important cue for disambiguation [6]. In practice, we adopt features derived from convolutional networks that have been trained for the task of object recognition. This is motivated by the versatility of these features, which have been successfully applied to many visual recognition tasks. For the network implementation, we adopt the *AlexNet* [18] architecture trained on ILSVRC-2012 [32] as provided by Caffe [15]. We retain activations of the network associated with the top-3 fully-connected layers (fc-6, fc-7, and fc-8). We encode temporal information in the activity by averaging activations across several frames. In practice, we compute these deep features every ten frames for all the videos in our dataset.

4.2. ActivityNet Benchmarks

We define three different application scenarios in which ActivityNet can be used for benchmarking. First, we investigate the performance of an activity recognition algorithm on the task of *Untrimmed video classification*. For the second task, we use the manually annotated trimmed video instances to construct the largest dataset for *Trimmed activity classification*. Finally, we benchmark *Activity detection* on all the untrimmed videos in ActivityNet.

4.2.1 Untrimmed Video Classification

In this task, we evaluate the capability of predicting activities in untrimmed video sequences. Here, videos can contain more than one activity, and typically large time lapses of the video are not related with any activity of interest.

Dataset: Using the labeled untrimmed videos from ActivityNet, we define a dataset for benchmarking untrimmed video classification. The dataset consists of 27801 videos that belong to 203 activity classes. We randomly split the data into three different subsets: train, validation and test, where 50% is used for training, and 25% for validation and testing.

Classifiers: Using the training set, we learn a set of linear SVM activity classifiers using a one-vs-all strategy. We use the validation set in order to tune the parameters of each classifier. Finally, we evaluate the models in the testing set, where the activity of each test video is predicted to be the one corresponding to the classifier with the largest margin.

Results: In this experiment, we measure the mean average precision (mAP) obtained by each activity classifier.

	Untrimmed Classification (mAP)		Trimmed Classification (mAP)	
Feature	**Validation**	**Test**	**Validation**	**Test**
Motion features (MF)				
HOG	29.2%	28.6%	35.9%	36.1%
HOF	32.7%	31.8%	40.1%	40.2%
MBH	34.1%	33.6%	41.7%	41.9%
Deep features (DF)				
fc-6	28.3%	28.1%	42.7%	43.1%
fc-7	28.0%	27.9%	41.1%	41.6%
fc-8	25.3%	24.9%	38.1%	38.2%
Per feature type				
MF	39.8%	39.2%	47.8%	47.6%
DF	28.9%	28.7%	43.7%	43.0%
SF	24.7%	24.5%	38.3%	37.9%
Combined				
MF+DF	**41.2%**	**40.9%**	**49.5%**	**49.1%**
MF+SF	**40.3%**	**40.1%**	**48.9%**	**48.6%**
DF+SF	32.7%	32.6%	44.2%	44.0%
MF+DF+SF	**42.5%**	**42.2%**	**50.5%**	**50.2%**

Table 1. Summary of classification results. The first two columns report results on the untrimmed video classification task, while the last two report results on trimmed video classification. The evaluation measure is mean average precision (mAP). We report validation and test performance, when different feature combinations are used. MF and DF refers to the concatenation of HOG, HOF and MBH features, and fc-6 and fc-7 respectively.

Since each untrimmed video may contain more than one activity label, we measure performance using mAP instead of a confusion matrix. Table 1 summarizes our results. We see that combining multiple features improves overall performance. Also, note that deep features obtain a competitive performance compared to the state-of-the-art improved trajectories features. The best results of deep features is obtained when we concatenate the activation of fc-6 and fc-7.

4.2.2 Trimmed Activity Classification

This task aims to predict the correct label of a trimmed video clip that contains a *single* activity instance. Here, we use all the trimmed activity instances annotated in ActivityNet to train classifiers and evaluate performance.

Dataset: We define a dataset for benchmarking human activity classification algorithms. The dataset includes 203 activity classes with 193 samples per category on average. These samples correspond to trimmed activity instances in ActivityNet. When generating the training, validation, and test subsets, we constrain the instances from a single video to be in the same subset so as to avoid data contamination.

Classifiers: As compared to untrimmed video classification, we build classifiers here with features that are *only* extracted from the trimmed activity itself. We learn a linear SVM classifier for each feature type. When combining multiple features, we simply sum the kernels before the learning procedure. To enable multi-class classification, we utilize a one-vs-all learning approach. Given a test video clip, we select the class with highest score.

Results: To measure recognition performance for this

task, we compute the mean average precision (mAP) over all the classes. As shown in Table 1, performance improves when multiple feature types are combined. As in the untrimmed video classification task, the DF model achieves a mAP score of 43.0% on the test subset. It reveals that these deep features by themselves encode discriminative information for human activities. We attribute this to the intuition that these features encode object appearance information and many activity categories involve human-object interactions.

4.2.3 Activity Detection

In this task, the goal is to find and recognize all activity instances within an untrimmed test video sequence. Activity detection algorithms should provide start and end frames, designating the duration of each activity present in the video. To evaluate the different classification models, we exploit ActivityNet annotations for the evaluation, thus, forming the largest and most diverse activity detection dataset in the literature.

Dataset: To the best of our knowledge, the ActivityNet-based detection dataset we use here is the largest existing dataset for this task. It contains a total of 849 hours of video, where 68.8 hours of video contain 203 human-centric activities. Here, we split the dataset in three different subsets as in the video classification tasks above.

Classifiers: We initialize our SVM models using the classifiers learned in the trimmed activity classification task. Then, we employ five rounds of hard negative mining, which generate a set of negative samples for each activity class. After each round, we only keep the hardest negatives in order to maintain a reasonable runtime. Given a test video sequence, we apply the learned classifiers using a sliding temporal window approach. From the training videos, we find that 7 temporal window lengths typically exist: 25, 60, 78, 100, 150, 190 and 250 frames. We then fix a sliding step size of 10 frames. Finally, we perform non-maximum suppression to ignore overlapping detection windows.

Results: To measure the performance of our model, we compute the mAP score over all activity classes. To do this, a detection is determined to be a true positive according to the following procedure: 1) we compute the overlap (measured by the intersection over union score) between a predicted temporal segment and a ground truth segment, 2) we mark the detection as positive if the overlap is greater than a threshold α. In practice, we vary the threshold α between 0.1 and 0.5. Table 2 summarizes the detection results. We see that MF consistently outperforms both DF and SF, across the different α values. In spite of the low performance of DF and SF, our model reveals a significant increase in performance when all feature types are combined. In general, it is clear that the detection task is very challenging for state-of-the-art detection methods.

Feature	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
MF	11.7%	11.4%	10.6%	9.7%	8.9%
DF	7.2%	6.8%	4.9%	4.1%	3.7%
SF	4.2%	3.9%	3.1%	2.1%	1.9%
MF+DF+SF	**12.5%**	**11.9%**	**11.1%**	**10.4%**	**9.7%**

Table 2. Summary of activity detection results. We report the mAP score for all activity classes. Due to the ambiguity inherent to the temporal annotation of activities, we use multiple values for the overlap threshold (α). We also investigate the performance of the different feature types, individually and collectively.

4.3. Discussions

We provide further analysis along three directions.

Qualitative results: Figure 5 shows some example results for the easiest and hardest activity classes for the tasks of untrimmed video classification and trimmed activity classification. These results are obtained using all three feature types. A sample set of correct detections are shown in the third column, while some hard false positive/negative samples for each activity class are shown in the last column. For untrimmed video classification, we find that the two easiest classes correspond to the *Sports and exercise* category. These activity classes are easier to classify, since they typically contain a repetitive and structured temporal sequence and are usually performed in similar scene contexts. We notice that activities occupying almost the entire video (in temporal length) are the hardest to classify. Regarding trimmed activity classification, the best classifiers tend to generate false positives when there are similar motions in the video. For example, the most confident false positives for the *Platform diving* class are from activities such as *Bungee jumping* and *Balance beam*, which contain motions that resemble those in platform diving. We also note that the most difficult classes tend to be confused with activities that have similar object or context appearance.

Where in the hierarchy are the easiest and hardest activity classes? To answer this question, we compute the mAP score per top level category. Table 3 shows these results for trimmed activity classification. We note that the activities related with *Sports and exercises* achieve the highest mAP. In contrast, *Household activities* achieve the lowest performance, due primarily to their unstructured nature, variability, and lack of temporal constraints. In contrast, *Sports and exercises* generally have a defined temporal ordering, and involve specific human-object interactions.

Comparing performance with existing datasets: To emphasize the difficulty of ActivityNet, we compare results for several datasets in Table 4. We consistently observe that ActivityNet constitutes a significant challenge to state-of-the-art recognition methods and is substantially more difficult than existing activity benchmarks. We attribute this to the following: a) ActivityNet increases the number of categories by a factor of two, and b) the variety in the video data represents a real world challenge for existing algorithms.

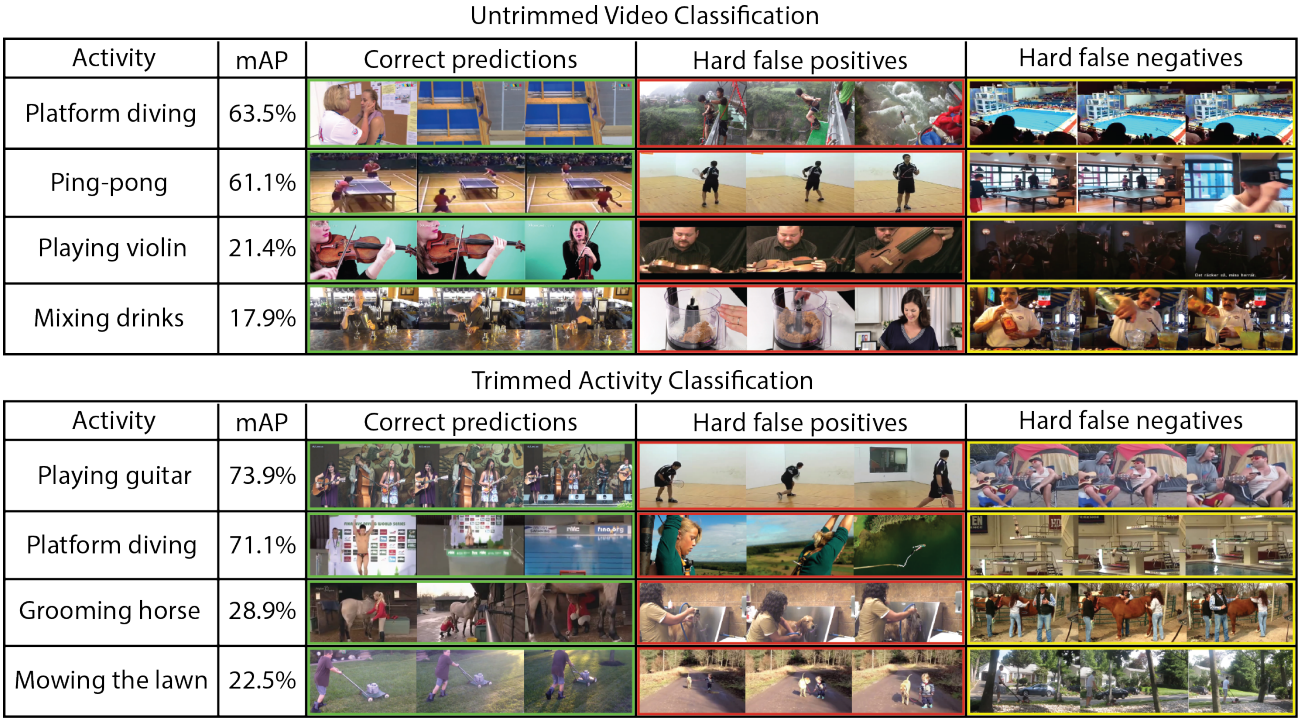

Figure 5. Example results for the two hardest and easiest activity classes in the untrimmed and trimmed classification tasks. Results are obtained using all three feature types (MF, DF, and SF). The third column shows some correct prediction samples for each class. The last two columns illustrate some hard false positive and hard false negative samples.

Category	Validation	Test
Household	34.2%	33.9%
Caring and helping	36.2%	36.7%
Personal care	41.5%	41.3%
Work-related	53.6%	53.1%
Eating and drinking	57.6%	57.2%
Socializing and leisure	63.8%	63.3%
Sports and exercises	66.6%	66.1%
Average	50.5%	50.2%

Table 3. Accuracy analysis on activity classification. We report mAP results for classifying each top-level class in ActivityNet. Here, all three feature types are used: motion, deep and static features.

Dataset	Method	Performance
Untrimmed video classification		
Thumos'14	[14]	71% (mAP)
Sports-1M	[16]	63.9% (mAP)
ActivityNet		42.2% (mAP)
Trimmed activity classification		
UCF101	[40]	85.9% (Accuracy)
HMDB51	[27]	66.7% (Accuracy)
ActivityNet		45.9% (Accuracy)
Activity detection		
Thumos'14	[25]	33.6% (mAP)
ActivityNet		11.9% (mAP)

Table 4. Cross-dataset performance comparison. State-of-the-art results are reported for each dataset. Reported results for the activity detection task corresponds to the performance obtained with $\alpha = 0.2$

5. Conclusions

In this paper, we introduce ActivityNet, a new large scale benchmark for human activity understanding. It is made possible by a large and continuous video collection and annotation effort that is easily scalable to larger numbers of activities and larger samples per activity, at a reasonably low cost. We compare ActivityNet with existing datasets for action/activity recognition. We show that ActivityNet presents more variety in terms of activity diversity and richness of taxonomy. It also contains more categories and samples per category than traditional action datasets. We also introduce three possible applications for using ActivityNet: untrimmed video classification, trimmed activity classification, and activity detection. The results obtained in these tasks reveal that ActivityNet unveils new challenges in understanding and recognizing human activities.

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, algorithmic baselines and a toolkit will be available at our website `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.

References

[1] B. Ainsworth, W. Haskell, S. Herrmann, N. Meckes, D. Bassett Jr., C. Tudor-Locke, J. Greer, J. Vezina, M. Whitt-Glover, and A. Leon. 2011 compendium of physical activities: a second update of codes and met values. *Medicine and Science in Sports and Exercise*, 2011.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014.

[3] I. Atmosukarto, B. Ghanem, and N. Ahuja. Trajectory-based fisher kernel representation for action recognition in videos. In *ICPR*. IEEE, 2012.

[4] I. Atmosukarto, B. Ghanem, and N. Ahuja. Action recognition using discriminative structured trajectory groups. 2015.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[6] V. Escorcia and J. C. Niebles. Spatio-temporal human-object interactions for action recognition in videos. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, 2013.

[7] FrameNet. `http://framenet.icsi.berkeley.edu`.

[8] Freebase: A community-curated database of well-known people, places, and things. `https://www.freebase.com`.

[9] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

[10] G. Guerra-Filho and Y. Aloimonos. Towards a sensorimotor WordNet SM: Closing the semantic gap. In *International WordNet Conference*, 2006.

[11] F. C. Heilbron and J. C. Niebles. Collecting and annotating human activities in web videos. In *International Conference on Multimedia Retrieval*, 2014.

[12] F. C. Heilbron, A. Thabet, J. C. Niebles, and B. Ghanem. Camera motion and surrounding scene appearance as context for action recognition. In *ACCV*, 2014.

[13] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In *ECCV*, 2010.

[14] M. Jain, J. van Gemert, and C. Snoek. University of amsterdam at THUMOS Challenge 2014. *ECCV THUMOS Challenge*, 2014.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, 2014.

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[17] A. R. Z. Khurram Soomro and M. Shah. A dataset of 101 human action classes from videos in the wild. Technical report, University of Central Florida, 2012.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

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

[20] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[21] I. Lillo, A. Soto, and J. C. Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *CVPR*, 2014.

[22] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[23] G. A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38:39–41, 1995.

[24] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[25] D. Oneata, J. Verbeek, and C. Schmid. The LEAR submission at Thumos 2014. In *ECCV THUMOS Challenge*, 2014.

[26] K. Pastra and Y. Aloimonos. The minimalist grammar of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1585):103–117, 2012.

[27] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *ECCV*, 2014.

[28] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[29] L. Pishchulin, M. Andriluka, and B. Schiele. Fine-grained activity recognition with holistic and pose based features. In X. Jiang, J. Hornegger, and R. Koch, editors, *Pattern Recognition*, volume 8753 of *Lecture Notes in Computer Science*, pages 678–689. Springer International Publishing, 2014.

[30] M. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[31] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[32] 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, 2014.

[33] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.

[34] K. D. Tang, B. Yao, L. Fei-Fei, and D. Koller. Combining the right features for complex event recognition. In *ICCV*, 2013.

[35] Thumos challenge 2014. `http://crcv.ucf.edu/THUMOS14`, 2013.

[36] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

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

[38] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013.

[39] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013.

[40] H. Wang and C. Schmid. Lear-inria submission for the thumos workshop. In *ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013.

[41] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu. Concurrent action detection with structural prediction. In *ICCV*, 2013.

[42] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. SUN Database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 2014.

[43] YouTube statistics. `http://www.youtube.com/yt/press/statistics.html`, 2015.