Moments in Time Dataset: one million videos for event understanding

Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, Aude Oliva

Abstract—We present the Moments in Time Dataset, a large-scale human-annotated collection of one million short videos corresponding to dynamic events unfolding within three seconds. Modeling the spatial-audio-temporal dynamics even for actions occurring in 3 second videos poses many challenges: meaningful events do not include only people, but also objects, animals, and natural phenomena; visual and auditory events can be symmetrical or not in time ("opening" means "closing" in reverse order), and transient or sustained. We describe the annotation process of our dataset (each video is tagged with one action or activity label among 339 different classes), analyze its scale and diversity in comparison to other large-scale video datasets for action recognition, and report results of several baseline models addressing separately and jointly three modalities: spatial, temporal and auditory. The Moments in Time dataset designed to have a large coverage and diversity of events in both visual and auditory modalities, can serve as a new challenge to develop models that scale to the level of complexity and abstract reasoning that a human processes on a daily basis.

Index Terms—video dataset, action recognition, event recognition

1 INTRODUCTION

"The best things in life are not things, they are moments" of raining, walking, splashing, resting, laughing, crying, jumping, etc. Moments happening in the world can unfold at time scales from a second to minutes, occur in different places, and involve people, animals, objects, as well as natural phenomena, like rain, wind, or just silence. Of particular interest are moments of a few seconds; they represent an ecosystem of changes in our surroundings that convey enough temporal information to interpret the auditory and visual dynamic world.

Here, we introduce the Moments in Time Dataset, a collection of one million short videos with a label each, corresponding to actions and events unfolding within 3 seconds.\(^1\) Crucially, temporal events of such length correspond to the average duration of human working memory [1], [6]. Working memory is a short-term memory-in-action buffer: it is specialized in representing information that is changing over time. Three seconds is a temporal envelope which holds meaningful actions between people, objects and phenomena (e.g. wind blowing, object falling on the floor, picking up something) or between actors (e.g. greeting someone, shaking hands, playing with a pet, etc).

 Bundling three seconds actions together allows for the creation of "compound" activities occurring at a longer time scale. For example, picking up an object, and carrying it away while running could be interpreted as the compound action "stealing", or "saving" or "delivering" depending on the social context of ownership and the type of place the activity occurs in. Hypothetically, when describing such a "stealing" event, one can go into the details of the movement of each joint and limb of the persons involved. However, this is not how we naturally describe compound events. Instead, we use verbs such as "picking", "carrying" and "running". These are the actions, which typically occur in a time window of 1-3 seconds. The ability to automatically recognize these short actions is a core step for automatic video comprehension.

The increasing availability of very large datasets (on the order of millions of labeled samples) is enabling rapid progress on challenging computer vision problems such as event and activity detection, common-sense interpretation or prediction of future events. Modeling the spatial-temporal dynamics even for events occurring in 3 second videos, poses a daunting challenge. For instance, inspecting videos in the dataset labeled with the action "opening", one can find people opening doors, gates, drawers, curtains and presents, animals and humans opening eyes, mouths and arms, and even a flower opening its petals. Furthermore, in some cases the same set of frames in reverse order actually depict a different action ("closing"). The temporal aspect in this case is crucial to recognition. Humans recognize that all of the above mentioned scenarios belong to the category "opening" even though visually they look very different from each other. There is a common transformation that occurs in space and time involving certain agents and/or objects that allows humans to associate it with the semantic meaning of the action "opening". The challenge is to develop models that recognize these transformations in a way that will allow them to discriminate between different actions, yet generalize to other agents and settings within the same action.

We expect the Moments in Time Dataset, the first version of which we present here, to enable models to richly

\(^1\) The website is http://moments.csail.mit.edu

\(^{1}\) The website is http://moments.csail.mit.edu
Fig. 1: Sample Videos. Day-to-day events can happen to many types of actors, in different environments, and at different scales. Moments in Time dataset has a significant intra-class variation among the categories. Here we illustrate one frame for a few video samples and actions. For example, car engines can open, books can open, and tulips can open.
understand actions and dynamics in videos. To the best of our knowledge, the collection is one of the largest human-annotated video datasets capturing visual and/or audible short events, produced by humans, animals, objects or nature. The classes are chosen such that they include the most commonly used verbs in the English language, covering a wide and diverse semantic space.

This work presents the first version of the Moments in Time dataset which includes one action label per video, and 339 different action classes. Clearly, there could be more than one action taking place even in a video that is three seconds long. This may hinder the performance of action recognition models which may predict an action correctly yet be penalized because the ground truth does not include that action. We therefore believe that the top 5 accuracy measure, commonly used in computer vision models to report classification performances, will be more meaningful for this version of the dataset. While the main purpose of this paper is to introduce the Moments in Time Dataset itself, in Section 4 we report experimental results of several known models trained and tested on the dataset, addressing separately and jointly three modalities: spatial, temporal and auditory.

As it is likely unfeasible to teach an exhaustive list of possible human-object-element interactions and activities, one strategy is to provide deep learning algorithms with a large coverage of the ecosystem of visual and auditory moments. The diversity of the Moments in Time dataset may enable models to learn discriminant information that is not necessarily taught in a fully supervised manner, allowing models to be more robust to unexpected events and generalize to novel situations and tasks.

2 RELATED WORK

Video Datasets: Large scale image datasets such as ImageNet [37] and Places [54], [55], have allowed great progress to be made for visual recognition in static images. Over the years, the size of video datasets for video understanding has grown steadily. The KTH [40] and Weizmann [8] were early datasets for human action understanding. Hollywood2 [32] used feature length movie films, and LabelMe video [50] used consumer video to create video datasets for action recognition and future prediction. The UCF101 dataset [41] and THUMOS [25] datasets are built from web videos that have become important benchmarks for video classification. JHMDB [24] has human activity categories with joints annotated. Kinetics [27] and YouTube-8M [2] introduced a large number of event categories by leveraging public videos from YouTube. The micro-videos dataset [33] uses social media videos to study an open-world vocabulary for video understanding. ActivityNet [9] explores recognizing activities in video and AVA [17] explores recognizing fine-grained actions with localization. The “something something” dataset [16] has crowdsourced workers collect a compositional video dataset, and Charades [42] uses crowdsourced workers to perform activities to collect video data. The VLOG dataset [14] and ADL [35] uses daily human activities to collect data with natural spatio-temporal context. As described below, two key features of the Moments in Time dataset are diversity and scale. In particular, we focus on brief moments where the agents are not limited to humans (for example, many objects can “fall” or “open”, see Figure 1).

Video Classification: The availability of video datasets has enabled significant progress at video understanding and classification. In early work, Laptev and Lindeberg [31] developed space-time interest point descriptors and Klaser et al. [29] designed histogram features for video. Pioneering work by Wang et al. [48] developed dense action trajectories by separating foreground motion from camera motion. Sadanand and Corso [38] designed ActionBank as a high-level representation for video and action classification, and Pirsiavash and Ramanan [36] leverage grammar models for temporally segmenting actions from video. Advances in deep convolutional networks have enabled large-scale video classification models [26], [43], [12], [47], [49], [10]. Various approaches of fusing RGB frames over the temporal dimension are explored on the Sport1M dataset [26]. Two stream CNNs with one stream of static images and the other stream of optical flows are proposed to fuse the information of object appearance and short-term motions [45]. 3D convolutional networks [47] use 3D convolution kernels to extract features from a sequence of dense RGB frames. Temporal Segment Networks sample frames and optical flow on different time segments to extract information for activity recognition [49]. A CNN+LSTM model, which uses a CNN to extract frame features and an LSTM to integrate features over time, is also used to recognize activities in videos [12]. Recently, 3D networks [10] use two stream CNNs with inflated 3D convolutions on both dense RGB and optical flow sequences to achieve state of the art performance on the Kinetics dataset [27].

Sound Classification: Environmental and ambient sound recognition is a rapidly growing area of research. Stowell et al. [45] collected an early dataset and assembled a challenge for sound classification, Piczak [34] collected a dataset of fifty sound categories and enough to train deep convolutional models, Salamon et al. [39] released a dataset of urban sounds, and Gemmeke et al. [15] use web videos for sound dataset collection. Recent work is now developing models for sound classification with deep neural networks. For example, Piczak [34] pioneered early work for convolutional networks for sound classification, Aytar et al. [4] transfer visual models into sound for auditory analysis, and Hershey et al. [20] develop large-scale convolutional models for sound classification, and Arandjelovic and Zisserman [3] train sound and vision representations jointly. In Moments in Time dataset, many videos have both visual and auditory signals, enabling for multi-modal video recognition.

3 THE MOMENTS IN TIME DATASET

The goal of this project is to design a high-coverage, high-density, balanced dataset of hundreds of verbs depicting moments of a few seconds. High-quality datasets should have broad coverage of the data space, high diversity and density of samples, and the ability to scale up. At the time this article is written, the first version of the Moments in Time Dataset consists of over 1,000,000 3-second videos corresponding to 339 different verbs depicting an action or activity. Each verb is associated with over 1,000 videos resulting in a large balanced dataset for learning a basis of
When building a large-scale dataset it is important to use an American English (COCA) [11]). We then clustered the verbs in order of most common to least common “washing, showering, bathing, soaping, grooming, shampooing, manicuring, moisturizing, and flossing”. Verbs can belong to multiple clusters due to their different frames of use. For instance, “washing” also belongs to a group associated with cleaning, mopping, scrubbing, etc.

Given these clusters, we then iteratively selected the most common verb from the most common cluster and added it to our vocabulary. The verb was then removed from all of its member clusters, and we repeated the process with the remaining verbs in the set. This method creates a list of verbs ordered according to not just the frequency of use of the verb, but also the frequency of its semantic meaning. From this sorted list of 4,500 verbs we then hand picked the 339 most common verbs that could be recognized in a 3-second video.

### 3.1 Building a Vocabulary of Active Moments

When building a large-scale dataset it is important to use an appropriate class vocabulary that contains a large coverage and diversity of classes. In order to ensure that we captured this criteria we began building our vocabulary by using the 4,500 most commonly used verbs from VerbNet [41] (according to the word frequencies in the Corpus of Contemporary American English (COCA) [11]). We then clustered the verbs according to their conceptual structure and meaning using the features for each verb from Propbank [28], FrameNet [5] and OntoNotes [21]. The clusters are sorted according to the combined frequency of use of each verb member of the cluster according to COCA. For example, we found a cluster associated with “grooming” which contained the following verbs in order of most common to least common “washing, showering, bathing, soaping, grooming, shampooing, manicuring, moisturizing, and flossing”.

### 3.2 Collection and Annotation

Once we form our vocabulary of the dataset, we crawl the internet and download videos related to each verb from a variety of different sources. This includes parsing video metadata and crawling search engines to build a list of candidate videos for each verb in our vocabulary. We randomly cut a 3-second section of each video and grouped the cut with the corresponding verb. These verb-video tuples are sent to Amazon Mechanical Turk for annotation.

Each worker is presented with the video-verb pair and asked to press a Yes or No key responding if the action is happening in the scene. The positive responses from the first round are then sent to a second round of annotation. Each HIT (one assignment for each worker) contains 64 different 3-second videos that are related to a single verb and 10 ground truth videos that are used for control. In each HIT, the first 4 questions are used to train the workers on the task and do not allow them to continue without selecting the correct answer. Only the results from HITs that earn a 90% or above on the control videos are included in the dataset. We chose this binary-classification setup because we have a large number of verb categories which makes class selection a difficult task for workers. We run each video in the training set through annotation at least 3 times and a require a human consensus of at least 75% to be considered a positive label. For the validation and test set we increase the minimum number of rounds of annotation to 4 with a human consensus of at least 85%. We do not set the threshold at 100% to allow for some videos that have actions that are slightly more difficult to recognize into the dataset. Figure 2 shows an example of the annotation task presented to the workers.

### 3.3 Dataset Statistics

A motivation for this project was to gather a large balanced and diverse dataset for training models for video understanding. Since we pull our videos from over 10 different sources we are able to include a large breadth of diversity that would be challenging using a single source. In total, we have collected over 1,000,000 labelled videos for 339 Moment classes. The graph on the left of Figure 3 shows the full distribution across all classes where the average number of labeled videos per class is 1,757 with a median of 2,775.

To further aid in building a diverse dataset we do not restrict the active agent in our videos to humans. Many events such as “walking”, “swimming”, “jumping”, and “carrying” are not specific to human agents. In addition, some classes may contain very few videos with human agents (e.g. “howling” or “flying”). True video understanding models should be able to recognize the event across agent classes. With this in mind we decided to build our dataset to be general across agents and present a new challenge to the field of video understanding. The middle graph in Figure 3 shows the distribution of the videos according to agent type (human, animal, object) for each class. On the
far left (larger human proportion), we have classes such as "typing", "sketching", and "repairing", while on the far right (smaller human proportion) we have events such as "storming", "roaring", and "erupting".

Another feature of the Moments in Time dataset is that we include sound-dependant classes. We do not restrict our videos to events that can be seen, if there is a moment that can only be heard in the video (e.g. "clapping" in the background) then we still include it. This presents another challenge in that purely visual models will not be sufficient to completely solve the dataset. The right graph in Figure 3 shows the distribution of videos according to whether or not the event in the video can be seen.

3.4 Dataset Comparisons

In order to highlight the key points of our dataset, we compare the scale, object-scene coverage, and the object-scene-action correlations found in Moments in Time to other large-scale video datasets for action recognition. These include UCF-101 [44], ActivityNet [9], Kinetics [27], Something-Something [16], AVA [17], and Charades [42]. Figure 4 compares the total number of action labels used for training (left) and the average number of videos per class in the training set (middle). This increase in scale for action recognition is beneficial for training large generalizable systems for machine learning.

Additionally, we compared the coverage of objects and scenes that can be recognized within the videos. This type of comparison helps to showcase the visual diversity of our dataset. To accomplish this, we extract 3 frames from each video evenly spaced at 25%, 50%, and 75% of the video duration and run a 50 layer resnet [18] trained on ImageNet [30] and a 50 layer resnet trained on Places [54] over each frame and average the prediction results for each video. We then compare the total number of objects and scenes recognized (top 1) by the networks in Figure 4 (right). The graph shows that 100% of the scene categories in Places and 99.9% of the object categories in ImageNet were recognized in our dataset. The closest dataset to ours in this comparison is Kinetics which has a recognized coverage of 99.5% of the scene categories in Places and 96.6% of the object categories in ImageNet. We should note that we are comparing the recognized categories from the top 1 prediction of each network. We have not annotated the scene locations and objects in each video of each dataset. However, a comparison of the visual features recognized by each network does still serve as an informative comparison of visual diversity.

4 EXPERIMENTS

In this section we present the details of our experimental setup utilized to obtain the reported baseline results.
we resize the RGB frames to a standard 340x256 pixels. In
with between 500 and 5,000 videos per class for 339 different
TABLE 1: Classification Accuracy: We show Top-1 and Top-5
accuracy of the baseline models on the validation set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Modality</th>
<th>Top-1 (%)</th>
<th>Top-5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td>-</td>
<td>0.29</td>
<td>1.47</td>
</tr>
<tr>
<td>ResNet50-scratch</td>
<td>Spatial</td>
<td>23.65</td>
<td>46.73</td>
</tr>
<tr>
<td>ResNet50-Places</td>
<td>Spatial</td>
<td>26.44</td>
<td>50.56</td>
</tr>
<tr>
<td>ResNet50-ImageNet</td>
<td>Spatial</td>
<td>27.16</td>
<td>51.68</td>
</tr>
<tr>
<td>TSN-Spatial</td>
<td>Spatial</td>
<td>24.11</td>
<td>49.10</td>
</tr>
<tr>
<td>BNInception-Flow</td>
<td>Temporal</td>
<td>11.60</td>
<td>27.40</td>
</tr>
<tr>
<td>ResNet50-DyImg</td>
<td>Temporal</td>
<td>15.76</td>
<td>35.69</td>
</tr>
<tr>
<td>TSN-Flow</td>
<td>Temporal</td>
<td>15.71</td>
<td>34.65</td>
</tr>
<tr>
<td>SoundNet</td>
<td>Auditory</td>
<td>7.60</td>
<td>18.00</td>
</tr>
<tr>
<td>TSN-2stream</td>
<td>Spatial+Temporal</td>
<td>25.32</td>
<td>50.10</td>
</tr>
<tr>
<td>TRN-Multiscale</td>
<td>Spatial+Temporal</td>
<td>28.27</td>
<td>53.87</td>
</tr>
<tr>
<td>Ensemble (average)</td>
<td>S+T+A</td>
<td>30.40</td>
<td>55.94</td>
</tr>
<tr>
<td>Ensemble (SVM)</td>
<td>S+T+A</td>
<td>30.42</td>
<td>55.60</td>
</tr>
</tbody>
</table>

4.1 Experimental Setup

Data. For training and testing models for video classification
on our dataset, we generate a training set of 802,264 videos
with between 500 and 5,000 videos per class for 339 different
Moment classes. We evaluate performance on a validation
set of 33,900 videos which consists of 100 videos for each
of the 339 classes. We additionally withhold a test set of
67,800 videos consisting of 200 videos per class which will be
used to evaluate submissions for a future action recognition
challenge.

Preprocessing. We extract RGB frames from the videos
at 25 fps. Given that the videos are with variable resolution,
we resize the RGB frames to a standard 340×256 pixels. In
the interest of performance, we pre-compute optical flow on
consecutive frames using an off-the-shelf implementation
of TVL1 optical flow algorithm [51] from the OpenCV
toolbox [23]. This formulation allows for discontinuities in
the optical flow field and thus more robust to noise. For
fast computation, we discretize the values of optical flow
fields into integers, clip the displacement with a maximum
absolute value of 15 and scale the range as 0–255. The
x and y displacements fields of every optical flow frame
can then be stored as two grayscale images with reduced
storage consumption. To correct for camera motion, we
subtract the mean vector from each displacement field in
the stack. For video frames, we use random cropping for
data augmentation and we subtract the ImageNet mean from
images.

Evaluation metric. We use the top-1 accuracy and top-5
classification accuracy as the scoring metrics. Top-1 accuracy
indicates the percentage of testing videos for which the top
confident predicted label is correct. Top-5 accuracy indicates
the percentage of the testing videos for which the ground-
truth label is among the top 5 ranked predicted labels. Top-5
accuracy is appropriate for video classification as videos may
contain multiple actions within them (see Figure 5).

4.2 Baselines for Video Classification

Here, we present several baselines for video classification
on the Moments in Time dataset. We show results for three
modalities (spatial, temporal, and auditory), as well as for
recent video classification models such as Temporal Segment
Networks [49] and Temporal Relation Networks [52]. We
further explore combining models to improve recognition
accuracy. The details of the baseline models grouped by
different modalities are listed below.

Spatial modality. We experiment with a 50 layer resnet
[19] network for classification given RGB frames of videos. In
training, the input to the network are randomly selected RGB
frames for each video. In testing, we average the prediction
from 6 equi-distant frames. We train the networks with
weights trained from scratch as ResNet50-scratch, initialized

Auditory modality. While many actions can be rec-
ognized visually, sound contains complementary or even
mandatory information for recognition of particular cate-
gories, such as cheering or talking, as can be seen in Figure
3 (right). We use raw waveforms as the input modality and
follow the network architecture from SoundNet [4] with
the output layer changed to predict moment categories. We
finetune a model pre-trained on 2,000,000 unlabeled videos

Temporal modality. We report results from two temporal
modality models. First, following [43], we compute optical
flow between adjacent frames encoded in Cartesian coordi-
nates as displacements. We use optic flow images by stacking
together 5 consecutive frames to form a 10 channel image (the
x and y displacement channels of optical flow). We use the
BNInception [22] as the base model, by modifying the first
convolutional layer to accept 10 input channels instead of 3
as BNInception-Flow. Second, we compute dynamic images [7]
as a means of spatiotemporal encoding of videos. A dynamic
image summarizes the gist of a video clip in a single image.
Dynamic images represent a video as a ranking function of
its frames using rankSVM [13]. RankSVM uses an implicit
video label - the frame ordering. We use a residual network
[19] with 50 layers as the architecture for training on the
dynamic images as ResNet50-DyImg.

We also train two recent action recognition models:
Temporal Segment Networks (TSN) [49] and Temporal
Relation Networks [52]. Temporal Segment Networks aim
to efficiently capture the long-range temporal structure of
videos using a sparse frame-sampling strategy. The TSN’s
spatial stream TSN-Spatial is fused with an optical flow
stream TSN-Flow via average consensus to form the two
stream TSN TSN-2stream. The base model for each stream is
a BNInception [22] model with three time segments.

Temporal Relation Networks (TRN) [52] are designed to
explicitly learn the temporal dependencies between video
segments that best characterize a particular action. This
“plug-and-play” module can model several short-range and
long-range temporal dependencies simultaneously to classify
actions that unfold at multiple time scales. In this paper, a
TRN with multi-scale relations TRN-Multiscale is trained
on the RGB frames only using InceptionV3 [46] as the
base model. The number of multi-scale relations used in
TRN is 8. Note that we classify the TRN-Multiscale as
spatiotemporal modality because in training it utilizes the
temporal dependency of different frames.

Ensemble. To combine different modalities for action
prediction, we conduct model ensemble over the top per-
foming model of each modality (spatial: ResNet50-ImageNet,
spatiotemporal: TRN-Multiscale, auditory: SoundNet). We try
two ensemble strategies: the first is average ensemble, in
Fig. 5: Overview of top detections for several single stream models. The ground truth label and top three model predictions are listed for representative frames of videos.
which we simply average the predicted class probability from each model; the second is SVM ensemble: we concatenate predicted class probabilities from each stream then fit a multi-class one-versus-all linear SVM to predict the moment categories (among a vocabulary of 339 verbs). SVM ensemble enables us to learn a weighted average of the modalities dependent on the category.

4.3 Baseline Results

Table 1 shows the Top-1 and Top-5 accuracy of the baseline models on the validation set. The best single model is the TRN-Multiscale, with a Top-1 accuracy of 28.27% and a Top-5 accuracy of 53.87%. The Ensemble model (average) gets the Top-5 accuracy as 55.94%.

Figure 5 illustrates some of the high scoring predictions from the baseline models. This qualitative result suggests that the models can recognize moments well when the action is well-framed and close up. However, the model frequently misfires when the category is fine-grained or there is background clutter. Figure 6 shows examples where the ground truth category is not detected in the top-5 predictions due to either significant background clutter or difficulty in recognizing actions across agents.

We visualize the prediction given by the model by generating the heatmaps for some video samples using the Class Activation Mapping (CAM) [53] in Figure 7. CAM highlights the most informative image regions relevant to the prediction. Here we use the top-1 prediction of the ResNet50-ImageNet model for each individual frame of the given video.

To understand some of the challenges, Figure 8 breaks down performance by category for different models and modalities. Categories that perform the best tend to have clear appearances and lower intra-class variation, for example bowling and surfing frequently happen in specific scene categories. The more difficult categories, such as covering, slipping, and plugging, tend to have wide spatiotemporal support as they can happen in most scenes and with most objects. Recognizing actions uncorrelated with scenes and objects seems to pose a challenge for video understanding.

Figure 8 also shows the roles that different modalities play in per category performance. Auditory models have a qualitatively different performance per category versus visual models, suggesting that sound provides a complementary signal to vision for recognizing actions in videos. However, the full ensemble model has per category performance that is fairly correlated with a single image, spatial model. Given the relatively low performance on Moments in Time, this suggests that there is still room to capitalize on temporal dynamics to better recognize action categories.

Figure 9 shows some of the most common confusions between categories. Generally, the most common failures are due to errors in fine-grained recognition, such as confusing submerging versus swimming, or lack of temporal reasoning, such as confusing opening versus closing. The confusions between a single frame model and the full model are qualitatively similar, suggesting that temporal reasoning remains a critical challenge for visual models. The auditory confusions, however, are qualitatively different, showing that sound is an important complementary signal for video understanding. We expect that, to advance performance in this dataset, models will need a rich understanding of dynamics, fine-grained recognition, and audio-visual reasoning.
Fig. 7: Predictions and Attention: We show some predictions (shown with class probability in top left corner) from ResNet50-ImageNet spatial model on held-out video data and the heatmaps which highlight the informative regions in some frames. For example, for recognizing the action chewing, the network focuses on the moving mouth.

5 CONCLUSION

We present the Moments in Time Dataset, a large-scale collection for video understanding, covering a wide class of dynamic events involving different agents (people, animals, objects, and natural phenomena), unfolding over three seconds. We report results of several baseline models addressing separately and jointly three modalities: spatial, temporal and auditory. This dataset presents a difficult task for the field of computer vision in that the labels correspond to different levels of abstraction (a verb like “falling” can apply to many different agents and scenarios, involving objects of different categories, see Figure 1). Thus it will serve as a new challenge to develop models that can appropriately scale to the level of complexity and abstract reasoning that a human processes on a daily basis.

Future versions of the dataset will include multi-labels action description (i.e. more than one action occurs in most 3-second videos, as illustrated in Figure 5), focus on growing the diversity of agents, and adding temporal transitions between the actions that agents performed. We also plan to organize challenges based on the various releases of the dataset, from general action recognition to more cognitive-level tasks such as modeling transformations and transfer learning across different agents and settings. For example, consider the challenge of training on actions performed solely by humans and testing on the same actions performed by animals. Humans are expert with analogies, the ability to seemingly transfer knowledge between events that have a partial similarity. At the core of common sense reasoning and creativity, analogies may occur across modalities, between different agents (a ball jumping to a person jumping) and at different levels of abstraction (e.g. opening a door and opening your hand to welcome someone). The project aims to produce datasets with a variety of levels of abstraction and agents (animate and inanimate agents performing similar actions), to serve as a step-stone towards the development of learning algorithms that are able to build analogies between things, imagine and synthesis novel events, and interpret compositional scenarios.

Acknowledgements: This work was supported by the MIT-IBM Watson AI Lab.

REFERENCES

Fig. 8: **Top 5 accuracy distribution per categories:** We show Left the best Ensemble model compared to SoundNet accuracy distribution. **Middle** shows the best Ensemble model compared to ResNet50-ImageNet, the best spatial only model. **Right** shows performances by SoundNet for the auditory modality. For visualization, only a subset of categories are labeled.

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.52</td>
<td>sailing</td>
<td>boating</td>
</tr>
<tr>
<td>0.41</td>
<td>grilling</td>
<td>barbecuing</td>
</tr>
<tr>
<td>0.37</td>
<td>emptying</td>
<td>filling</td>
</tr>
<tr>
<td>0.35</td>
<td>closing</td>
<td>opening</td>
</tr>
<tr>
<td>0.33</td>
<td>parading</td>
<td>marching</td>
</tr>
<tr>
<td>0.32</td>
<td>calling</td>
<td>telephoning</td>
</tr>
<tr>
<td>0.28</td>
<td>barking</td>
<td>howling</td>
</tr>
<tr>
<td>0.27</td>
<td>sketching</td>
<td>drawing</td>
</tr>
<tr>
<td>0.25</td>
<td>submerging</td>
<td>swimming</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.53</td>
<td>grilling</td>
<td>barbecuing</td>
</tr>
<tr>
<td>0.45</td>
<td>sailing</td>
<td>boating</td>
</tr>
<tr>
<td>0.37</td>
<td>emptying</td>
<td>filling</td>
</tr>
<tr>
<td>0.34</td>
<td>drawing</td>
<td>sketching</td>
</tr>
<tr>
<td>0.30</td>
<td>cooking</td>
<td></td>
</tr>
<tr>
<td>0.29</td>
<td>marching</td>
<td>parading</td>
</tr>
<tr>
<td>0.28</td>
<td>howling</td>
<td>calling</td>
</tr>
<tr>
<td>0.28</td>
<td>raining</td>
<td>barking</td>
</tr>
</tbody>
</table>

Fig. 9: **Most Confused Categories:** We show the most commonly confused categories by three models. **Left** shows confusions by the ensemble model that combines spatial, temporal, and auditory modalities. **Middle** shows confusions by the best spatial only model (ResNet50ImageNet) . **Right** shows confusions by the sound model. The first column of each table shows the frequency of the confusion.

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.21</td>
<td>barking</td>
<td>howling</td>
</tr>
<tr>
<td>0.19</td>
<td>adult female singing</td>
<td>child singing</td>
</tr>
<tr>
<td>0.15</td>
<td>sneezing</td>
<td>spitting</td>
</tr>
<tr>
<td>0.15</td>
<td>cheerleading</td>
<td>cheering</td>
</tr>
<tr>
<td>0.13</td>
<td>storming</td>
<td>raining</td>
</tr>
<tr>
<td>0.14</td>
<td>bowling</td>
<td>barking</td>
</tr>
<tr>
<td>0.13</td>
<td>child speaking</td>
<td>child singing</td>
</tr>
<tr>
<td>0.11</td>
<td>shouting</td>
<td>cheering</td>
</tr>
<tr>
<td>0.11</td>
<td>singing</td>
<td>child singing</td>
</tr>
<tr>
<td>0.11</td>
<td>protesting</td>
<td>shouting</td>
</tr>
</tbody>
</table>
