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 Gutfruend, 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

M Monfort, B Zhou, A Andonian, T Yan, K Ramakrishnan, C Vondrick, A Oliva are with Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139 USA.
L Brown, Q Fan, D Gutfreund are with International Business Machines, 75 Binney St., Cambridge, MA 02142 USA.
SA Bargal is with Boston University, 111 Cummington Mall, Boston, MA 02215 USA.

1. The website is http://moments.csail.mit.edu

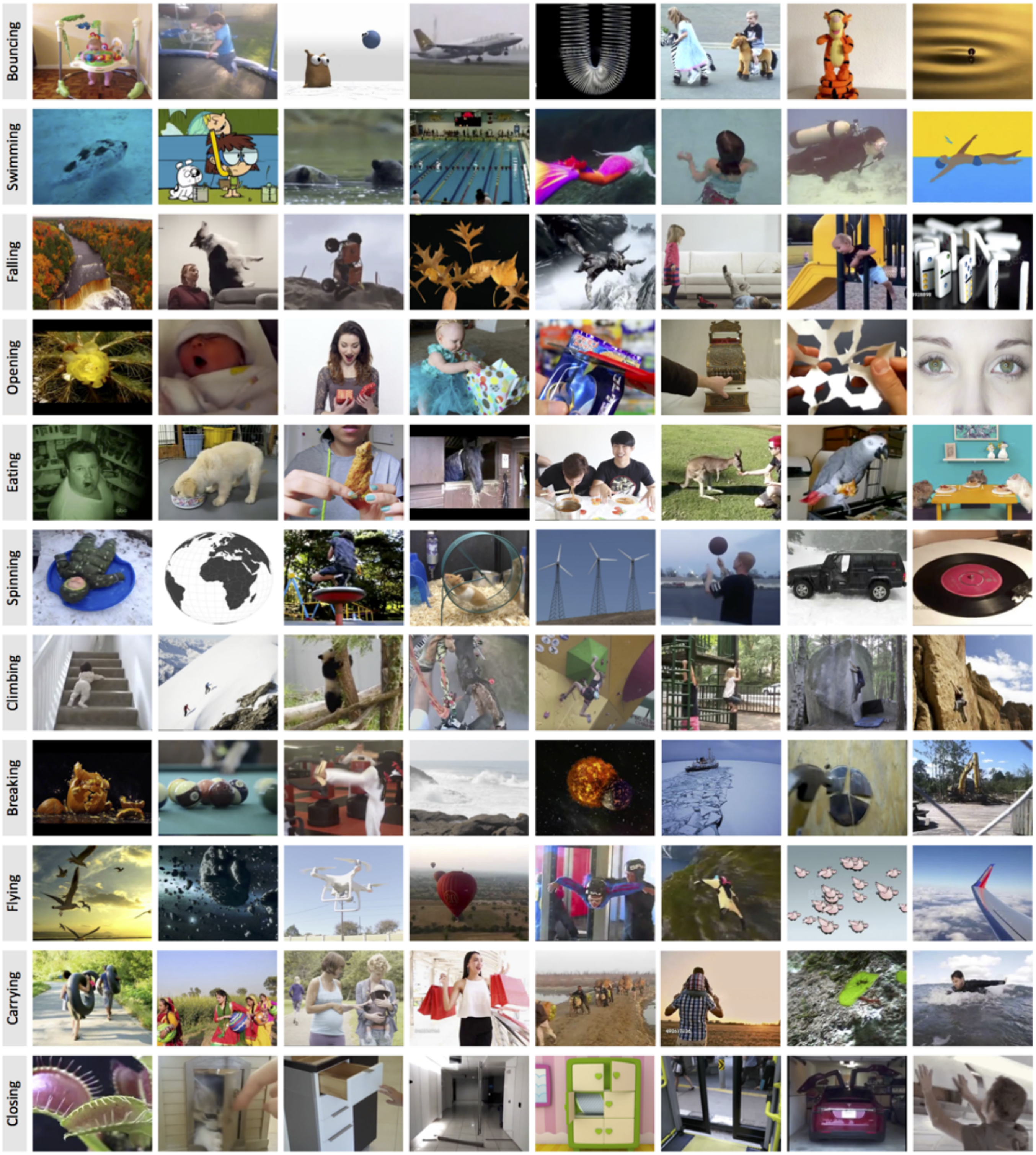

Fig. 1: **Sample Videos.** Day-to-day events can happen to many types of actors, in different environments, and at different scales. Moments in Time dataset has a significant intra-class variation among the categories. Here we illustrate one frame for a few video samples and actions. For example, car 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 [44] 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 seperating 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 [43]. 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, I3D 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, Piczack [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 Arandjelović 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 dynamical events from videos. Whereas several verbs and actions are often needed to describe the richness of a few seconds event (see examples in Figure 5), the first release, comes with one ground truth verb per video. Importantly, the dataset is designed to have, and to grow towards, a very large diversity of both inter-class and intra-class variation that represent a dynamical event at different levels of abstraction (i.e. "opening" doors, drawers, curtains, presents, eyes, mouths, even a flower opening its petals).

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". 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.

3.2 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 [2]. 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

2. Sources: Youtube (www.youtube.com), Flickr (www.flickr.com), Vine (www.vine.co), Metacafe (www.metacafe.com), Peeks (www.peeks.com), Vimeo (www.vimeo.com), VideoBlocks (www.videoblocks.com), Bing (www.bing.com), Giphy (www.giphy.com), The Weather Channel (www.weather.com) and Getty-Images (www.gettyimages.com)

Fig. 2: **User interface.** An example for our binary annotation task for the action cooking.

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

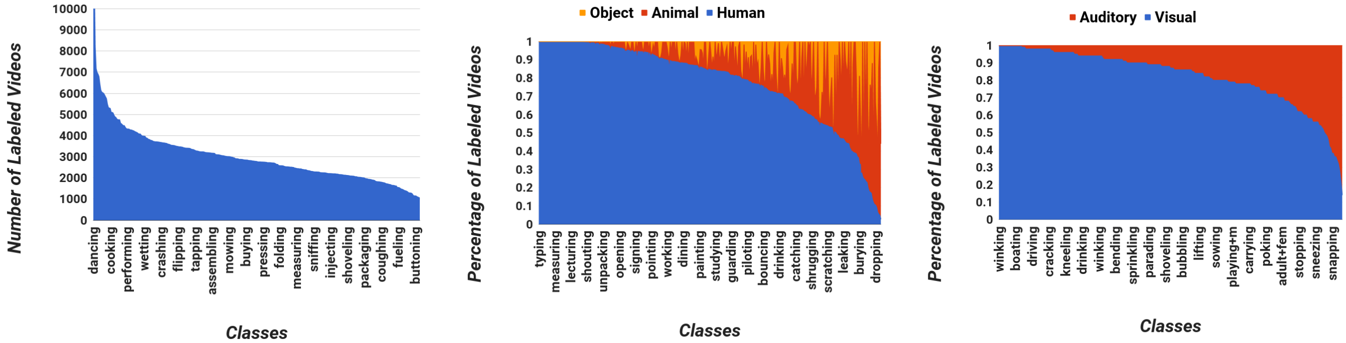

Fig. 3: **Dataset Statistics. Left**: Distribution of the number of videos belonging to each category. **Middle**: Per class distribution of videos that have humans, animals, or objects as agents completing actions. **Right**: Per class distribution of videos that require audio to recognize the class category and videos that can be categorized with only visual information.

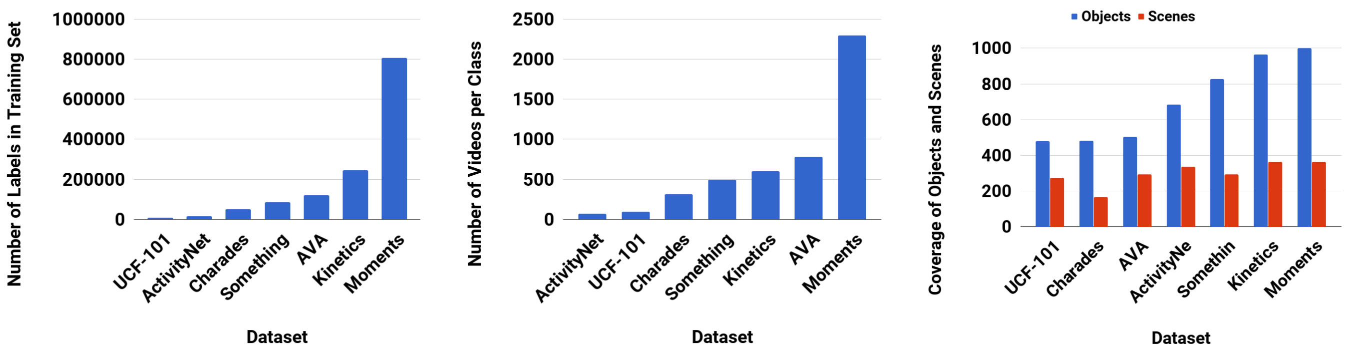

Fig. 4: **Comparison to Datasets.** For each dataset we provide different comparisons. **Left**: the total number of action labels in the training set. **Middle**: the average number of videos per class (some videos can belong to multiple classes).**Right**: the coverage of objects and scenes recognized (top 1) by networks trained on Places and Imagenet.

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 that belong to each 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.

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

TABLE 1: **Classification Accuracy:** We show Top-1 and Top-5 accuracy of the baseline models on the validation set.

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 340x256 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-scrach*, initialized on Places [55] as *ResNet50-Places*, and initialized on ImageNet [30] as *ResNet50-ImageNet*.

Auditory modality. While many actions can be recognized visually, sound contains complementary or even mandatory information for recognition of particular categories, 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 from Flickr [4] as *SoundNet*.

Temporal modality. We report results from two temporal modality models. First, following [43], we compute optical flow between adjacent frames encoded in Cartesian coordinates 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 performing model of each modality (spatial: *ResNet50-ImageNet*, spatiotemporal: *TRN-Multiscale*, auditory: *SoundNet*). We try two ensemble strategies: the first is average ensemble, in

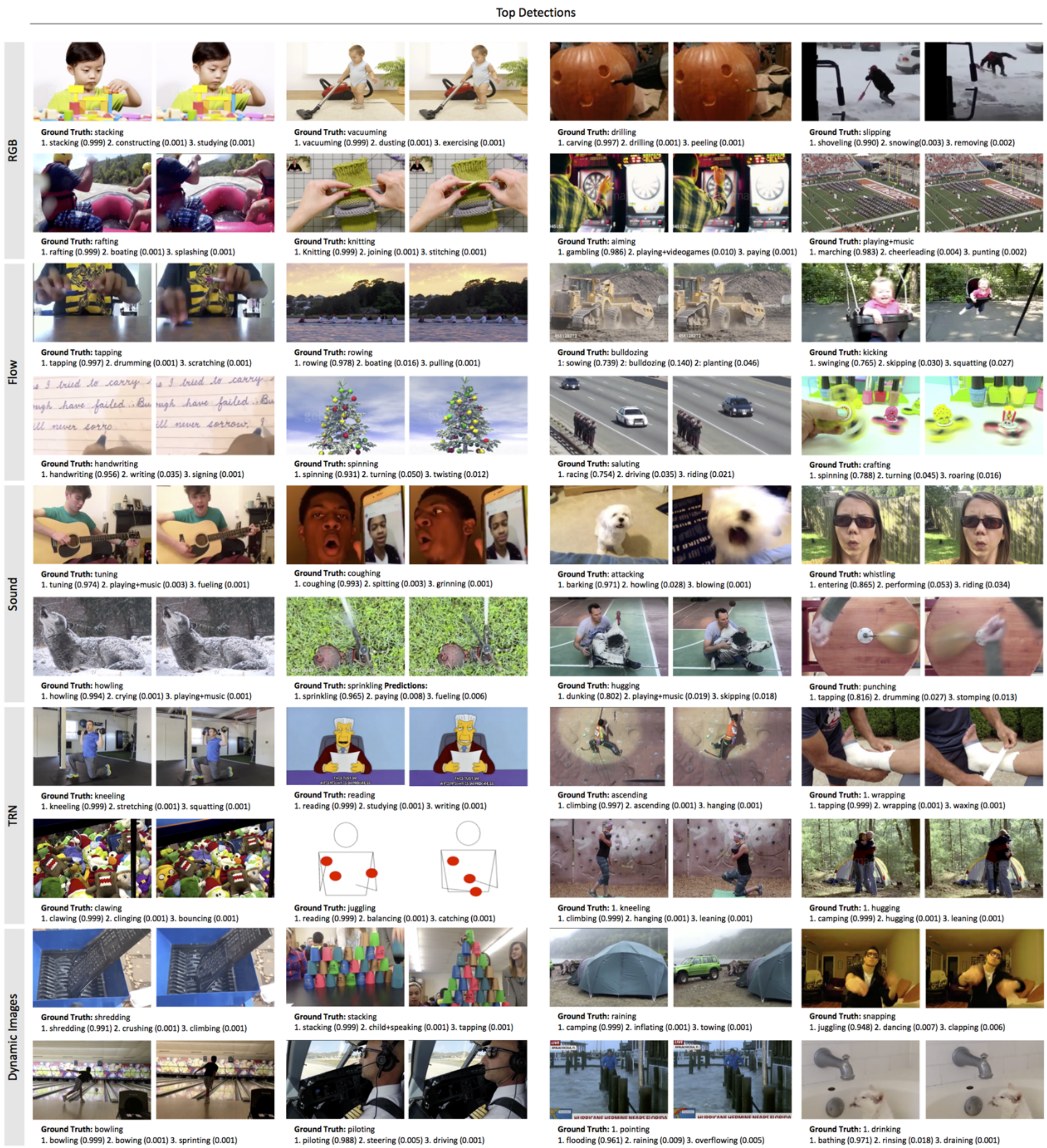

Fig. 5: **Overview of top detections for several single stream models.** The ground truth label and top three model predictions are listed for representative frames of videos.

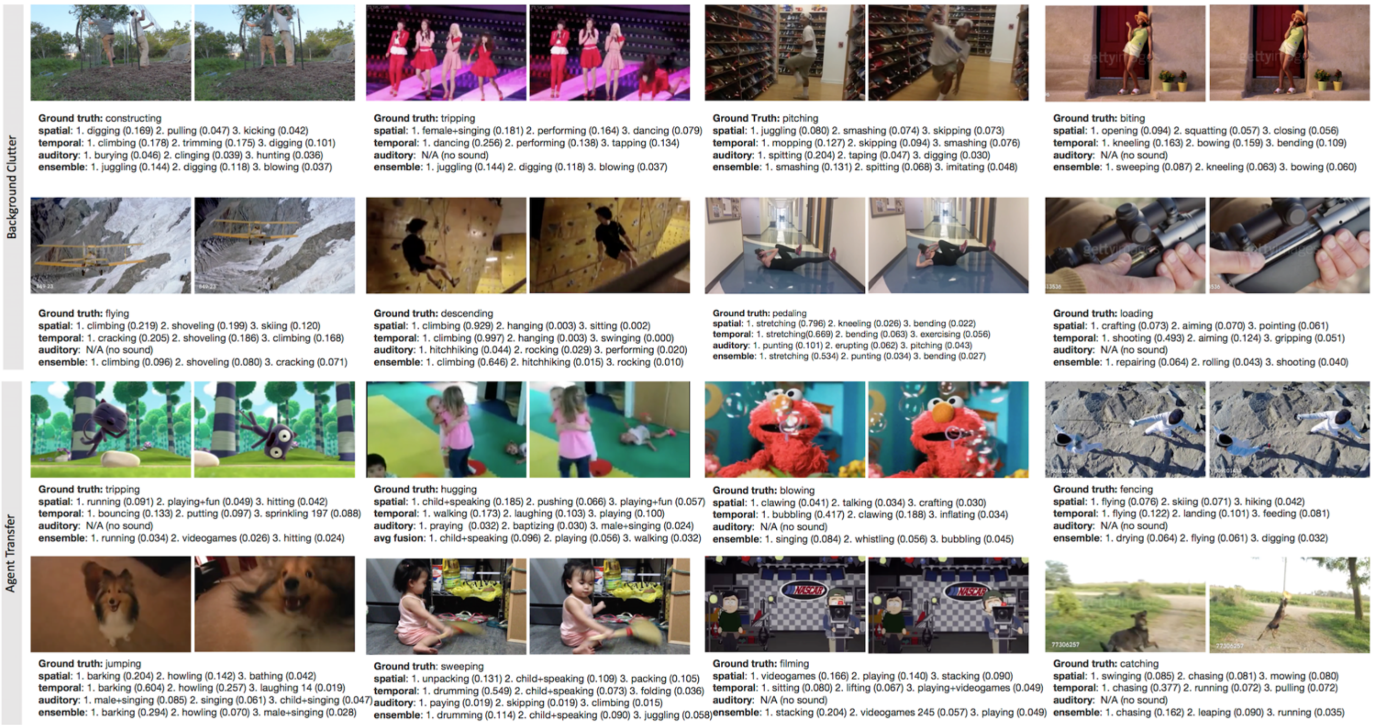

Fig. 6: **Examples of missed detections:** We show examples of videos where the prediction is not in the top-5. Common failures are often due to background clutter or poor generalization across agents (humans, animals, objects).

which we simply average the predicted class probability from each models; 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 qualtiatively 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.

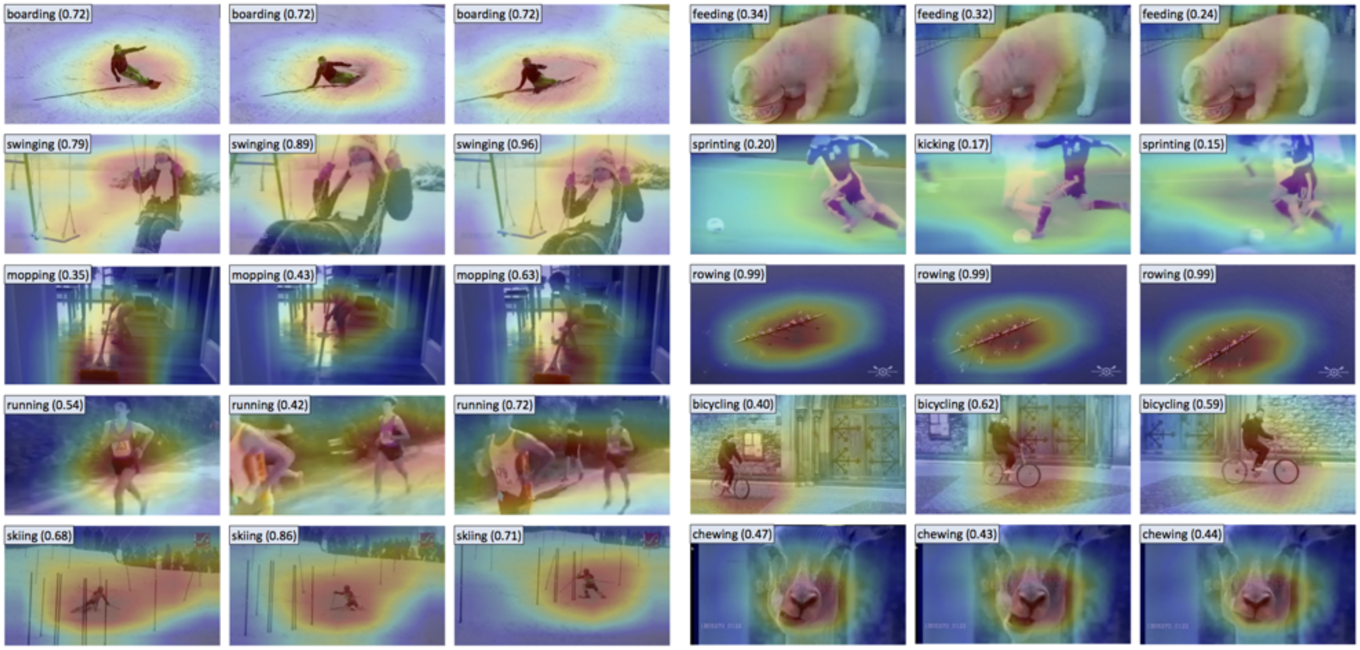

Fig. 7: **Predictions and Attention:** We show some predictions (shown with class probability in top left corner) from *ResNet50-ImageNet* spatial model on held-out video data and the heatmaps which highlight the informative regions in some frames. For example, for recognizing the action chewing, the 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

[1] Baddeley A. Working memory. *Science:255.556-559*, 1992.
[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.
[3] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. *arXiv preprint arXiv:1705.08168*, 2017.
[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 892–900. Curran Associates, Inc., 2016.
[5] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.
[6] P. Barrouillet, S. Bernardin, and V. Camos. Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General:133.83*, 2004.
[7] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
[8] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE, 2005.
[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

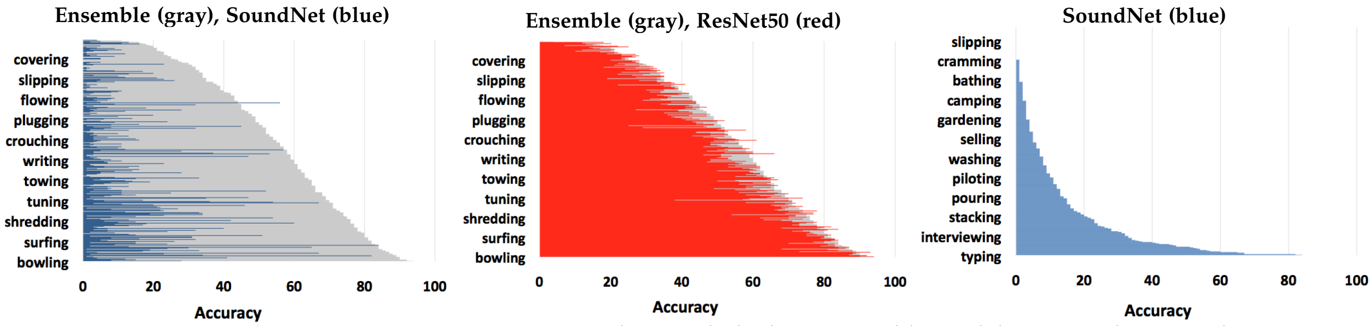

Fig. 8: **Top 5 accuracy distribution per categories**: We show **Left** the best Ensemble model compared to SoundNet accuracy distribution. **Middle** shows the best Ensemble model compared to ResNet50-ImageNet, the best spatial only model. **Right** shows performances by SoundNet for the auditory modality.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

Spatial, Temporal, Auditory (Ensemble)		
Freq.	Actual	Predicted
0.52	sailing	boating
0.41	grilling	barbecuing
0.37	emptying	filling
0.35	closing	opening
0.33	parading	marching
0.33	storming	raining
0.32	calling	telephoning
0.28	barking	howling
0.27	sketching	drawing
0.25	submerging	swimming

Spatial (ResNet50ImageNet)		
Freq.	Actual	Predicted
0.53	grilling	barbecuing
0.45	sailing	boating
0.37	closing	opening
0.34	emptying	filling
0.32	sketching	drawing
0.30	frying	cooking
0.29	parading	marching
0.28	calling	telephoning
0.28	barking	howling
0.28	storming	raining

Auditory (SoundNet)		
Freq.	Actual	Predicted
0.21	barking	howling
0.19	adult female singing	child singing
0.15	sneezing	spitting
0.15	cheerleading	cheering
0.15	storming	raining
0.14	howling	barking
0.13	child speaking	child singing
0.11	shouting	cheering
0.11	singing	child singing
0.11	protesting	shouting

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.

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *Proc. ICCV*, 2017.

[11] Mark Davies. Corpus of contemporary american english (coca), 2016.

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[13] Basura Fernando and Stephen Gould. Learning end-to-end video classification with rank-pooling. In *International Conference on Machine Learning*, pages 1187–1196, 2016.

[14] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. *arXiv preprint arXiv:1712.02310*, 2017.

[15] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dartaset for audio events. In *IEEE ICASSP*, 2017.

[16] Raghav Goyal, Samira Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. *arXiv preprint arXiv:1706.04261*, 2017.

[17] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. *arXiv preprint arXiv:1705.08421*, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. *arXiv preprint arXiv:1609.09430*, 2016.

[21] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: The 90 In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[23] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015.

[24] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, pages 3192–3199, December 2013.

[25] YG Jiang, J Liu, A Roshan Zamir, G Toderici, I Laptev, M Shah, and R Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014.

[26] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[28] Paul Kingsbury and Martha Palmer. In *Proceedings of Treebanks and Lexical Theories*, Växjö, Sweden.

[29] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[31] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *9th International Conference on Computer Vision, Nice, France*, pages 432–439. IEEE conference proceedings, 2003.

[32] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[33] Phuc Xuan Nguyen, Gregory Rogez, Charless Fowlkes, and Deva Ramanan. The open world of micro-videos. *arXiv preprint arXiv:1603.09439*, 2016.

[34] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pages 1–6. IEEE, 2015.

[35] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847–2854. IEEE, 2012.

[36] Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 612–619, 2014.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[38] Sreemananath Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012.

[39] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044. ACM, 2014.

[40] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[41] Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, Philadelphia, PA, USA, 2005. AAI3179808.

[42] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. *CoRR*, abs/1604.01753, 2016.

[43] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[44] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[45] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[48] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[49] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. ECCV*, 2016.

[50] Jenny Yuen, Bryan Russell, Ce Liu, and Antonio Torralba. Labelme video: Building a video database with human annotations. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1451–1458. IEEE, 2009.

[51] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. *Pattern Recognition*, pages 214–223, 2007.

[52] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. *arXiv preprint arXiv:1711.08496*, 2017.

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[54] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[55] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, 2014.