AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions

Chunhui Gu, Chen Sun, David Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul R Sukthankar, et al.

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions

Chunhui Gu* Chen Sun* David A. Ross* Carl Vondrick* Caroline Pantofaru*
Yeqing Li* Sudheendra Vijayanarasimhan* George Toderici* Susanna Ricco*
Rahul Sukthankar* Cordelia Schmid†* Jitendra Malik‡*

Abstract

This paper introduces a video dataset of spatio-temporally localized Atomic Visual Actions (AVA). The AVA dataset densely annotates 80 atomic visual actions in 437 15-minute video clips, where actions are localized in space and time, resulting in 1.59M action labels with multiple labels per person occurring frequently. The key characteristics of our dataset are: (1) the definition of atomic visual actions, rather than composite actions; (2) precise spatio-temporal annotations with possibly multiple annotations for each person; (3) exhaustive annotation of these atomic actions over 15-minute video clips; (4) people temporally linked across consecutive segments; and (5) using movies to gather a varied set of action representations. This departs from existing datasets for spatio-temporal action recognition, which typically provide sparse annotations for composite actions in short video clips.

AVA, with its realistic scene and action complexity, exposes the intrinsic difficulty of action recognition. To benchmark this, we present a novel approach for action localization that builds upon the current state-of-the-art methods, and demonstrates better performance on JHMDB and UCF101-24 categories. While setting a new state of the art on existing datasets, the overall results on AVA are low at 15.8% mAP, underscoring the need for developing new approaches for video understanding.

1. Introduction

We introduce a new annotated video dataset, AVA, to advance action recognition research (see Fig. 1). The annotation is person-centric at a sampling frequency of 1 Hz. Every person is localized using a bounding box and the attached labels correspond to (possibly multiple) actions being performed by the actor: one action corresponding to the actor's **pose** (orange text) — standing, sitting, walking, swimming etc. — and there may be additional actions corresponding to **interactions with objects** (red text) or **interactions with other persons** (blue text). Each person in a frame containing multiple actors is labeled separately.

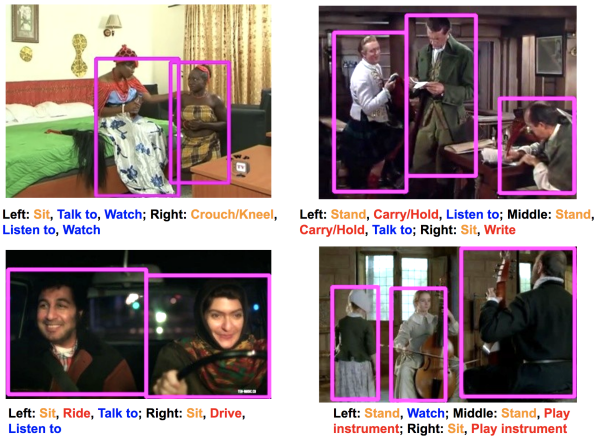

Figure 1. The bounding box and action annotations in sample frames of the AVA dataset. Each bounding box is associated with 1 pose action (in orange), 0–3 interactions with objects (in red), and 0–3 interactions with other people (in blue). Note that some of these actions require temporal context to accurately label.

To label the actions performed by a person, a key choice is the annotation vocabulary, which in turn is determined by the temporal granularity at which actions are classified. We use short segments (± 1.5 seconds centered on a keyframe) to provide temporal context for labeling the actions in the middle frame. This enables the annotator to use movement cues for disambiguating actions such as pick up or put down that cannot be resolved in a static frame. We keep the temporal context relatively brief because we are interested in (temporally) fine-scale annotation of physical actions, which motivates "Atomic Visual Actions" (AVA). The vocabulary consists of 80 different atomic visual actions. Our dataset is sourced from the 15th to 30th minute time intervals of 437 different movies, which given the 1 Hz sampling frequency gives us 900 keyframes for each movie. In each keyframe, every person is labeled with (possibly multiple) actions from the AVA vocabulary. Each person is linked to the consecutive keyframes to provide short temporal sequences of action labels (Section 4.3). We now motivate the

*Google Research
†Inria, Laboratoire Jean Kuntzmann, Grenoble, France
‡University of California at Berkeley, USA

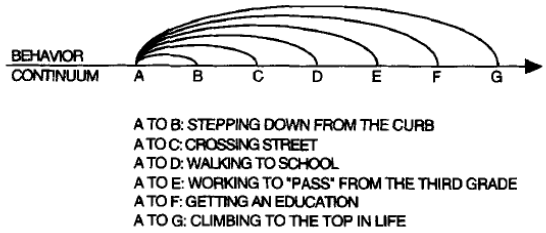

Figure 2. This figure illustrates the hierarchical nature of an activity. From Barker and Wright [3], pg. 247.

main design choices of AVA.

Atomic action categories. Barker & Wright [3] noted the hierarchical nature of activity (Fig. 2) in their classic study of the "behavior episodes" in the daily lives of the residents of a small town in Kansas. At the finest level, the actions consist of atomic body movements or object manipulation but at coarser levels, the most natural descriptions are in terms of intentionality and goal-directed behavior.

This hierarchy makes defining a vocabulary of action labels ill posed, contributing to the slower progress of our field compared to object recognition; exhaustively listing high-level behavioral episodes is impractical. However if we limit ourselves to fine time scales, then the actions are very physical in nature and have clear visual signatures. Here, we annotate keyframes at 1 Hz as this is sufficiently dense to capture the complete semantic content of actions while enabling us to avoid requiring unrealistically precise temporal annotation of action boundaries. The THUMOS challenge [18] observed that action boundaries (unlike objects) are inherently fuzzy, leading to significant inter-annotator disagreement. By contrast, annotators can easily determine (using $\pm$1.5s of context) whether a frame *contains* a given action. Effectively, AVA localizes action start and end points to an acceptable precision of ± 0.5 s.

Person-centric action time series. While events such as trees falling do not involve people, our focus is on the activities of people, treated as single agents. There could be multiple people as in sports or two people hugging, but each one is an agent with individual choices, so we treat each separately. The action labels assigned to a person over time is a rich source of data for temporal modeling (Section 4.3).

Annotation of movies. Ideally we would want behavior "in the wild". We do not have that, but movies are a compelling approximation, particularly when we consider the diversity of genres and countries with flourishing film industries. We do expect some bias in this process. Stories have to be interesting and there is a grammar of the film language [2] that communicates through the juxtaposition of shots. That said, in each shot we can expect an unfolding sequence of human actions, somewhat representative of reality, as conveyed by competent actors. AVA complements the current datasets sourced from user generated video because we expect movies to contain a greater range of activities as befits the telling of diverse stories.

Exhaustive action labeling. We label all the actions of all the people in all the keyframes. This will naturally result in a Zipf's law type of imbalance across action categories. There will be many more examples of typical actions (standing or sitting) than memorable ones (dancing), but this is how it should be! Recognition models need to operate on realistic "long tailed" action distributions [15] rather than being scaffolded using artificially balanced datasets. Another consequence of our protocol is that since we do not retrieve examples of action categories by explicit querying of internet video resources, we avoid a certain kind of bias: opening a door is a common event that occurs frequently in movie clips; however a door opening action that has been tagged as such on YouTube is likely attention worthy in a way that makes it atypical.

We believe that AVA, with its realistic complexity, exposes the inherent difficulty of action recognition hidden by many popular datasets in the field. A video clip of a single person performing a visually salient action like swimming in typical background is easy to discriminate from, say, one of a person running. Compare with AVA where we encounter multiple actors, small in image size, performing actions which are only subtly different such as touching vs. holding an object. To verify this intuition, we do comparative bench-marking on JHMDB [20], UCF101-24 categories [32] and AVA. The approach we use for spatio-temporal action localization (see Section 5) builds upon multi-frame approaches [16, 41], but classifies tubelets with I3D convolutions [6]. We obtain state-of-the-art performance on JHMDB [20] and UCF101-24 categories [32] (see Section 6) while the mAP on AVA is only 15.8%.

The AVA dataset has been released publicly at https://research.google.com/ava/.

2. Related work

Action recognition datasets. Most popular action classification datasets, such as KTH [35], Weizmann [4], Hollywood-2 [26], HMDB [24], UCF101 [39] consist of short clips, manually trimmed to capture a single action. These datasets are ideally suited for training fully-supervised, whole-clip, forced-choice video classifiers. Recently, datasets, such as TrecVid MED [29], Sports-1M [21], YouTube-8M [1], Something-something [12], SLAC [48], Moments in Time [28], and Kinetics [22] have focused on large-scale video classification, often with automatically generated – and hence potentially noisy – annotations. They serve a valuable purpose but address a different need than AVA.

Some recent work has moved towards temporal localization. ActivityNet [5], THUMOS [18], MultiTHUMOS [46] and Charades [37] use large numbers of untrimmed videos,

each containing multiple actions, obtained either from YouTube (ActivityNet, THUMOS, MultiTHUMOS) or from crowdsourced actors (Charades). The datasets provide temporal (but not spatial) localization for each action of interest. AVA differs from them, as we provide spatio-temporal annotations for each subject performing an action and annotations are dense over 15-minute clips.

A few datasets, such as CMU [23], MSR Actions [47], UCF Sports [32] and JHMDB [20] provide spatio-temporal annotations in each frame for short videos. The main differences with our AVA dataset are: the small number of actions; the small number of video clips; and the fact that clips are very short. Furthermore, actions are composite (e.g., pole-vaulting) and not atomic as in AVA. Recent extensions, such as UCF101 [39], DALY [44] and Hollywood2Tubes [27] evaluate spatio-temporal localization in untrimmed videos, which makes the task significantly harder and results in a performance drop. However, the action vocabulary is still restricted to a limited number of composite actions. Moreover, they do not densely cover the actions; a good example is BasketballDunk in UCF101, where only the dunking player is annotated. However, real-world applications often require a continuous annotations of atomic actions of all humans, which can then be composed into higher-level events. This motivates AVA's exhaustive labeling over 15-minute clips.

AVA is also related to still image action recognition datasets [7, 9, 13] that are limited in two ways. First, the lack of motion can make action disambiguation difficult. Second, modeling composite events as a *sequence* of atomic actions is not possible in still images. This is arguably out of scope here, but clearly required in many real-world applications, for which AVA does provide training data.

Methods for spatio-temporal action localization. Most recent approaches [11, 30, 34, 43] rely on object detectors trained to discriminate action classes at the frame level with a two-stream variant, processing RGB and flow data separately. The resulting per-frame detections are then linked using dynamic programming [11, 38] or tracking [43]. All these approaches rely on integrating frame-level detections. Very recently, multi-frame approaches have emerged: Tubelets [41] jointly estimate localization and classification over several frames, T-CNN [16] use 3D convolutions to estimate short tubes, micro-tubes rely on two successive frames [33] and pose-guided 3D convolutions add pose to a two-stream approach [49]. We build upon the idea of spatio-temporal tubes, but employ state-of-the-art I3D convolution [6] and Faster R-CNN [31] region proposals to outperform the state of the art.

3. Data collection

Annotation of the AVA dataset consists of five stages: action vocabulary generation, movie and segment selection, person bounding box annotation, person linking and action annotation.

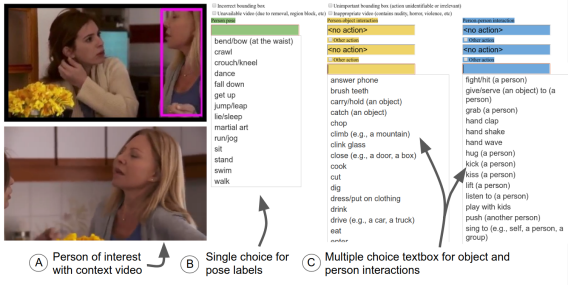

Figure 3. User interface for action annotation. Details in Sec 3.5.

3.1. Action vocabulary generation

We follow three principles to generate our action vocabulary. The first one is generality. We collect generic actions in daily-life scenes, as opposed to specific activities in specific environments (e.g., playing basketball on a basketball court). The second one is atomicity. Our action classes have clear visual signatures, and are typically independent of interacted objects (e.g., hold without specifying what object to hold). This keeps our list short yet complete. The last one is exhaustivity. We initialized our list using knowledge from previous datasets, and iterated the list in several rounds until it covered $\sim$99% of actions in the AVA dataset labeled by annotators. We end up with 14 pose classes, 49 person-object interaction classes and 17 person-person interaction classes in the vocabulary.

3.2. Movie and segment selection

The raw video content of the AVA dataset comes from YouTube. We begin by assembling a list of top actors of many different nationalities. For each name we issue a YouTube search query, retrieving up to 2000 results. We only include videos with the "film" or "television" topic annotation, a duration of over 30 minutes, at least 1 year since upload, and at least 1000 views. We further exclude black & white, low resolution, animated, cartoon, and gaming videos, as well as those containing mature content.

To create a representative dataset within constraints, our selection criteria avoids filtering by action keywords, using automated action classifiers, or forcing a uniform label distribution. We aim to create an international collection of films by sampling from large film industries. However, the depiction of action in film is biased, e.g. by gender [10], and does not reflect the "true" distribution of human activity.

Each movie contributes equally to the dataset, as we only label a sub-part ranging from the 15th to the 30th minute. We skip the beginning of the movie to avoid annotating titles or trailers. We choose a duration of 15 minutes so we are able to include more movies under a fixed annotation budget, and thus increase the diversity of our dataset.

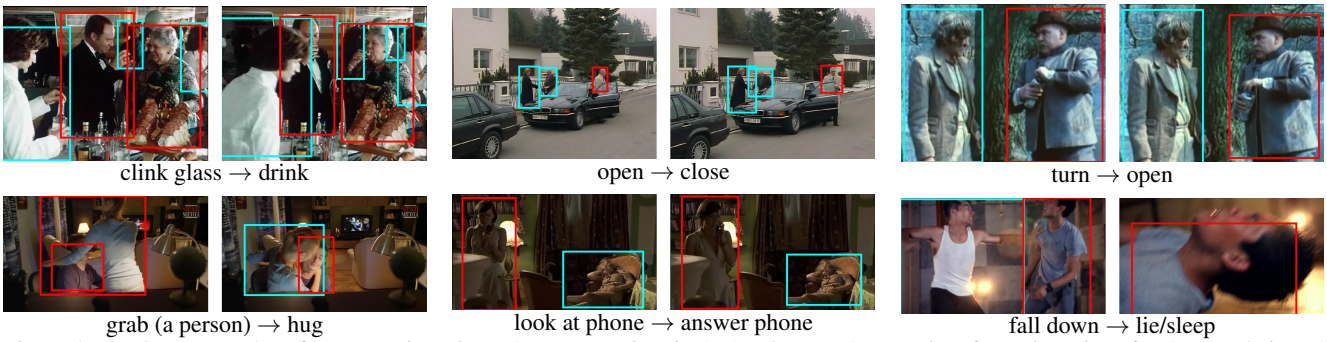

Figure 4. We show examples of how atomic actions change over time in AVA. The text shows pairs of atomic actions for the people in red bounding boxes. Temporal information is key for recognizing many of the actions and appearance can substantially vary within an action category, such as opening a door or bottle.

Each 15-min clip is then partitioned into 900 overlapping 3s movie segments with a stride of 1 second.

3.3. Person bounding box annotation

We localize a person and his or her actions with a bounding box. When multiple subjects are present in a keyframe, each subject is shown to the annotator separately for action annotation, and thus their action labels can be different.

Since bounding box annotation is manually intensive, we choose a hybrid approach. First, we generate an initial set of bounding boxes using the Faster-RCNN person detector [31]. We set the operating point to ensure high-precision. Annotators then annotate the remaining bounding boxes missed by our detector. This hybrid approach ensures full bounding box recall which is essential for benchmarking, while minimizing the cost of manual annotation. This manual annotation retrieves only 5% more bounding boxes missed by our person detector, validating our design choice. Any incorrect bounding boxes are marked and removed by annotators in the next stage of action annotation.

3.4. Person link annotation

We link the bounding boxes over short periods of time to obtain ground-truth person tracklets. We calculate the pairwise similarity between bounding boxes in adjacent key frames using a person embedding [45] and solve for the optimal matching with the Hungarian algorithm [25]. While automatic matching is generally strong, we further remove false positives with human annotators who verify each match. This procedure results in 81,000 tracklets ranging from a few seconds to a few minutes.

3.5. Action annotation

The action labels are generated by crowd-sourced annotators using the interface shown in Figure 3. The left panel shows both the middle frame of the target segment (top) and the segment as a looping embedded video (bottom). The bounding box overlaid on the middle frame specifies the person whose action needs to be labeled. On the right are text boxes for entering up to 7 action labels, including 1 pose action (required), 3 person-object interactions (optional), and 3 person-person interactions (optional). If none of the listed actions is descriptive, annotators can flag a check box called "other action". In addition, they could flag segments containing blocked or inappropriate content, or incorrect bounding boxes.

In practice, we observe that it is inevitable for annotators to miss correct actions when they are instructed to find all correct ones from a large vocabulary of 80 classes. Inspired by [36], we split the action annotation pipeline into two stages: action proposal and verification. We first ask multiple annotators to propose action candidates for each question, so the joint set possesses a higher recall than individual proposals. Annotators then verify these proposed candidates in the second stage. Results show significant recall improvement using this two-stage approach, especially on actions with fewer examples. See detailed analysis in the supplemental material. On average, annotators take 22 seconds to annotate a given video segment at the propose stage, and 19.7 seconds at the verify stage.

Each video clip is annotated by three independent annotators and we only regard an action label as ground truth if it is verified by at least two annotators. Annotators are shown segments in randomized order.

3.6. Training, validation and test sets

Our training/validation/test sets are split at the video level, so that all segments of one video appear only in one split. The 437 videos are split into 239 training, 64 validation and 134 test videos, roughly a 55:15:30 split, resulting in 215k training, 57k validation and 120k test segments.

4. Characteristics of the AVA dataset

We first build intuition on the diversity and difficulty of our AVA dataset through visual examples. Then, we characterize the annotations of our dataset quantitatively. Finally, we explore action and temporal structure.

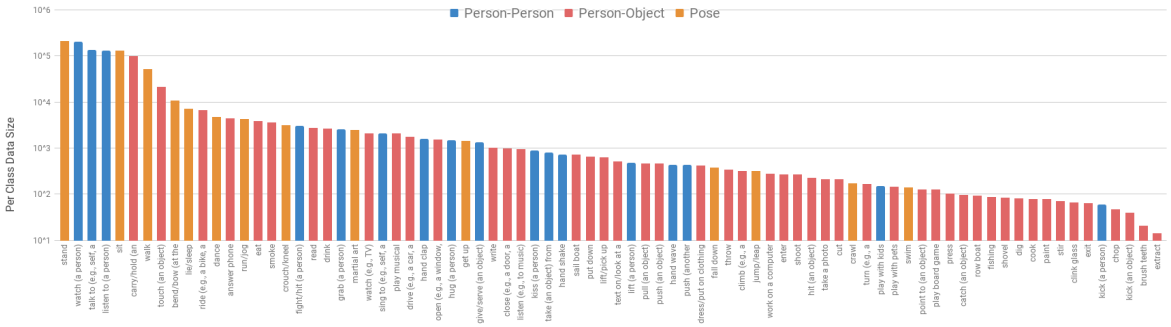

Figure 5. Sizes of each action class in the AVA train/val dataset sorted by descending order, with colors indicating action types.

4.1. Diversity and difficulty

Figure 4 shows examples of atomic actions as they change over consecutive segments. Besides variations in bounding box size and cinematography, many of the categories will require discriminating fine-grained differences, such as "clinking glass" versus "drinking" or leveraging temporal context, such as "opening" versus "closing".

Figure 4 also shows two examples for the action "open". Even within an action class the appearance varies with vastly different contexts: the object being opened may even change. The wide intra-class variety will allow us to learn features that identify the critical spatio-temporal parts of an action — such as the breaking of a seal for "opening".

4.2. Annotation Statistics

Figure 5 shows the distribution of action annotations in AVA. The distribution roughly follows Zipf's law. Figure 6 illustrates bounding box size distribution. A large portion of people take up the full height of the frame. However, there are still many boxes with smaller sizes. The variability can be explained by both zoom level as well as pose. For example, boxes with the label "enter" show the typical pedestrian aspect ratio of 1:2 with average widths of 30% of the image width, and an average heights of 72%. On the other hand, boxes labeled "lie/sleep" are close to square, with average widths of 58% and heights of 67%. The box widths are widely distributed, showing the variety of poses people undertake to execute the labeled actions.

There are multiple labels for the majority of person bounding boxes. All bounding boxes have one pose label, 28% of bounding boxes have at least 1 person-object interaction label, and 67% of them have at least 1 person-person interaction label.

4.3. Temporal Structure

A key characteristic of AVA is the rich temporal structure that evolves from segment to segment. Since we have linked people between segments, we can discover common consecutive actions by looking at pairs of actions performed by the same person. We sort pairs by Normalized Pointwise

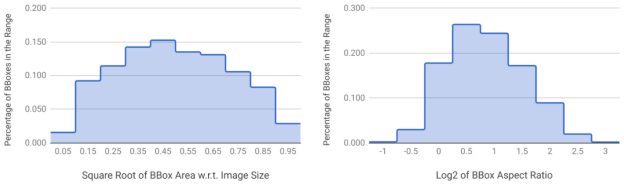

Figure 6. Size and aspect ratio variations of annotated bounding boxes in the AVA dataset. Note that our bounding boxes consist of a large variation of sizes, many of which are small and hard to detect. Large variation also applies to the aspect ratios of bounding boxes, with mode at 2:1 ratio (e.g., sitting pose).

Mutual Information (NPMI) [8], which is commonly used in linguistics to represent the co-occurrence between two words: $\mathrm{NPMI}(x,y) = \left(\ln \frac{p(x,y)}{p(x)p(y)}\right) / \left(-\ln p(x,y)\right)$. Values intuitively fall in the range $(-1, 1]$, with -1 for pairs of words that never co-occur, 0 for independent pairs, and 1 for pairs that always co-occur.

Table 1 shows pairs of actions with top NPMI in consecutive one-second segments for the same person. After removing identity transitions, some interesting common sense temporal patterns arise. Frequently, there are transitions from "look at phone" $\rightarrow$ "answer phone", "fall down" $\rightarrow$ "lie", or "listen to" $\rightarrow$ "talk to". We also analyze interperson action pairs. Table 2 shows top pairs of actions performed at the same time, but by different people. Several meaningful pairs emerge, such as "ride" $\leftrightarrow$ "drive", "play music" $\leftrightarrow$ "listen", or "take" $\leftrightarrow$ "give/serve". The transitions between atomic actions, despite the relatively coarse temporal sampling, provide excellent data for building more complex models of actions and activities with longer temporal structure.

5. Action Localization Model

Performance numbers on popular action recognition datasets such as UCF101 or JHMDB have gone up considerably in recent years, but we believe that this may present an artificially rosy picture of the state of the art. When the video clip involves only a single person performing something visually characteristic like swimming in an equally characteristic background scene, it is easy to classify ac-

First Action	Second Action	NPMI
ride (eg bike/car/horse)	drive (eg car/truck)	0.68
watch (eg TV)	work on a computer	0.64
drive (eg car/truck)	ride (eg car bike/car/horse)	0.63
open (eg window/door)	close (eg door/box)	0.59
text on/look at a cellphone	answer phone	0.53
listen to (person)	talk to (person)	0.47
fall down	lie/sleep	0.46
talk to (person)	listen to (person)	0.43
stand	sit	0.40
walk	stand	0.40

Table 1. We show top pairs of consecutive actions that are likely to happen before/after for the same person. We sort by NPMI.

Person 1 Action	Person 2 Action	NPMI
ride (eg bike/car/horse)	drive (eg car/truck)	0.60
play musical instrument	listen (eg music)	0.57
take (object)	give/serve (object)	0.51
talk to (person)	listen to (person)	0.46
stand	sit	0.31
play musical instrument	dance	0.23
walk	stand	0.21
watch (person)	write	0.15
walk	run/jog	0.15
fight/hit (a person)	stand	0.14

Table 2. We show top pairs of simultaneous actions by different people. We sort by NPMI.

curately. Difficulties come in when actors are multiple, or small in image size, or performing actions which are only subtly different, and when the background scenes are not enough to tell us what is going on. AVA has these aspects galore, and we will find that performance at AVA is much poorer as a result. Indeed this finding was foreshadowed by the poor performance at the Charades dataset [37].

To prove our point, we develop a state of the art action localization approach inspired by recent approaches for spatio-temporal action localization that operate on multi-frame temporal information [16, 41]. Here, we rely on the impact of larger temporal context based on I3D [6] for action detection. See Fig. 7 for an overview of our approach.

Following Peng and Schmid [30], we apply the Faster RCNN algorithm [31] for end-to-end localization and classification of actions. However, in their approach, the temporal information is lost at the first layer where input channels from multiple frames are concatenated over time. We propose to use the Inception 3D (I3D) architecture by Carreira and Zisserman [6] to model temporal context. The I3D architecture is designed based on the Inception architecture [40], but replaces 2D convolutions with 3D convolutions. Temporal information is kept throughout the network. I3D achieves state-of-the-art performance on a wide range of video classification benchmarks.

To use I3D with Faster RCNN, we make the following changes to the model: first, we feed input frames of length T to the I3D model, and extract 3D feature maps of

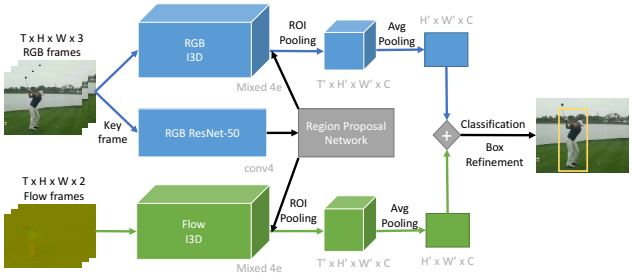

Figure 7. Illustration of our approach for spatio-temporal action localization. Region proposals are detected and regressed with Faster-RCNN on RGB keyframes. Spatio-temporal tubes are classified with two-stream I3D convolutions.

size $T' \times W' \times H' \times C$ at the *Mixed 4e* layer of the network. The output feature map at *Mixed 4e* has a stride of 16, which is equivalent to the conv4 block of ResNet [14]. Second, for action proposal generation, we use a 2D ResNet-50 model on the keyframe as the input for the region proposal network, avoiding the impact of I3D with different input lengths on the quality of generated action proposals. Finally, we extend ROI Pooling to 3D by applying the 2D ROI Pooling at the same spatial location over all time steps. To understand the impact of optical flow for action detection, we fuse the RGB stream and the optical flow stream at the feature map level using average pooling.

Baseline. To compare to a frame-based two-stream approach on AVA, we implement a variant of [30]. We use Faster RCNN [31] with ResNet-50 [14] to jointly learn action proposals and action labels. Region proposals are obtained with the RGB stream only. The region classifier takes as input RGB along with optical flow features stacked over 5 consecutive frames. As for our I3D approach, we jointly train the RGB and the optical flow streams by fusing the conv4 feature maps with average pooling.

Implementation details. We implement FlowNet v2 [19] to extract optical flow features. We train Faster-RCNN with asynchronous SGD. For all training tasks, we use a validation set to determine the number of training steps, which ranges from 600K to 1M iterations. We fix the input resolution to be 320 by 400 pixels. All the other model parameters are set based on the recommended values from [17], which were tuned for object detection. The ResNet-50 networks are initialized with ImageNet pre-trained models. For the optical flow stream, we duplicate the conv1 filters to input 5 frames. The I3D networks are initialized with Kinetics [22] pre-trained models, for both the RGB and optical flow streams. Note that although I3D were pre-trained on 64-frame inputs, the network is fully convolutional over time and can take any number of frames as input. All feature layers are jointly updated during training. The output frame-level detections are post-processed with non-maximum suppression with threshold 0.6.

One key difference between AVA and existing action detection datasets is that the action labels of AVA are not mu-

tually exclusive. To address this, we replace the standard softmax loss function by a sum of binary Sigmoid losses, one for each class. We use Sigmoid loss for AVA and softmax loss for all other datasets.

Linking. Once we have per frame-level detections, we link them to construct action tubes. We report video-level performance based on average scores over the obtained tubes. We use the same linking algorithm as described in [38], except that we do not apply temporal labeling. Since AVA is annotated at 1 Hz and each tube may have multiple labels, we modify the video-level evaluation protocol to estimate an upper bound. We use ground truth links to infer detection links, and when computing IoU score of a class between a ground truth tube and a detection tube, we only take tube segments that are labeled by that class into account.

6. Experiments and Analysis

We now experimentally analyze key characteristics of AVA and motivate challenges for action understanding.

6.1. Datasets and Metrics

AVA benchmark. Since the label distribution in AVA roughly follows Zipf's law (Figure 5) and evaluation on a very small number of examples could be unreliable, we use classes that have at least 25 instances in validation and test splits to benchmark performance. Our resulting benchmark consists of a total of 214,622 training, 57,472 validation and 120,332 test examples on 60 classes. Unless otherwise mentioned, we report results trained on the training set and evaluated on the validation set. We randomly select 10% of the training data for model parameter tuning.

Datasets. Besides AVA, we also analyze standard video datasets in order to compare difficulty. JHMDB [20] consists of 928 trimmed clips over 21 classes. We report results for split one in our ablation study, but results are averaged over three splits for comparison to the state of the art. For UCF101, we use spatio-temporal annotations for a 24-class subset with 3207 videos, provided by Singh *et al*. [38]. We conduct experiments on the official split1 as is standard.

Metrics. For evaluation, we follow standard practice when possible. We report intersection-over-union (IoU) performance on frame level and video level. For frame-level IoU, we follow the standard protocol used by the PASCAL VOC challenge [9] and report the average precision (AP) using an IoU threshold of 0.5. For each class, we compute the average precision and report the average over all classes. For video-level IoU, we compute 3D IoUs between ground truth tubes and linked detection tubes at the threshold of 0.5. The mean AP is computed by averaging over all classes.

6.2. Comparison to the state-of-the-art

Table 3 shows our model performance on two standard video datasets. Our 3D two-stream model obtains state-of-the-art performance on UCF101 and JHMDB, outperforming well-established baselines for both frame-mAP and video-mAP metrics.

Frame-mAP	JHMDB	UCF101-24
Actionness [42]	39.9%	-
Peng w/o MR [30]	56.9%	64.8%
Peng w/ MR [30]	58.5%	65.7%
ACT [41]	65.7%	69.5%
Our approach	**73.3%**	**76.3%**

Video-mAP	JHMDB	UCF101-24
Peng w/ MR [30]	73.1%	35.9%
Singh *et al*. [38]	72.0%	46.3%
ACT [41]	73.7%	51.4%
TCNN [16]	76.9%	-
Our approach	**78.6%**	**59.9%**

Table 3. Frame-mAP (top) and video-mAP (bottom) @ IoU 0.5 for JHMDB and UCF101-24. For JHMDB, we report averaged performance over three splits. Our approach outperforms previous state-of-the-art on both metrics by a considerable margin.

However, the picture is less auspicious when recognizing atomic actions. Table 4 shows that the same model obtains relatively low performance on AVA validation set (frame-mAP of **15.8%**, video-mAP of **12.3%** at 0.5 IoU and **17.9%** at 0.2 IoU), as well as test set (frame-mAP of **14.7%**). We attribute this to the design principles behind AVA: we collected a vocabulary where context and object cues are not as discriminative for action recognition. Instead, recognizing fine-grained details and rich temporal models may be needed to succeed at AVA, posing a new challenge for visual action recognition. In the remainder of this paper, we analyze what makes AVA challenging and discuss how to move forward.

6.3. Ablation study

How important is temporal information for recognizing AVA categories? Table 4 shows the impact of the temporal length and the type of model. All 3D models outperform the 2D baseline on JHMDB and UCF101-24. For AVA, 3D models perform better after using more than 10 frames. We can also see that increasing the length of the temporal window helps for the 3D two-stream models across all datasets. As expected, combining RGB and optical flow features improves the performance over a single input modality. Moreover, AVA benefits more from larger temporal context than JHMDB and UCF101, whose performances saturate at 20 frames. This gain and the consecutive actions in Table 1 suggests that one may obtain further gains by leveraging the rich temporal context in AVA.

How challenging is localization versus recognition? Table 5 compares the performance of end-to-end action localization and recognition versus class agnostic action localization. We can see that although action localization is more

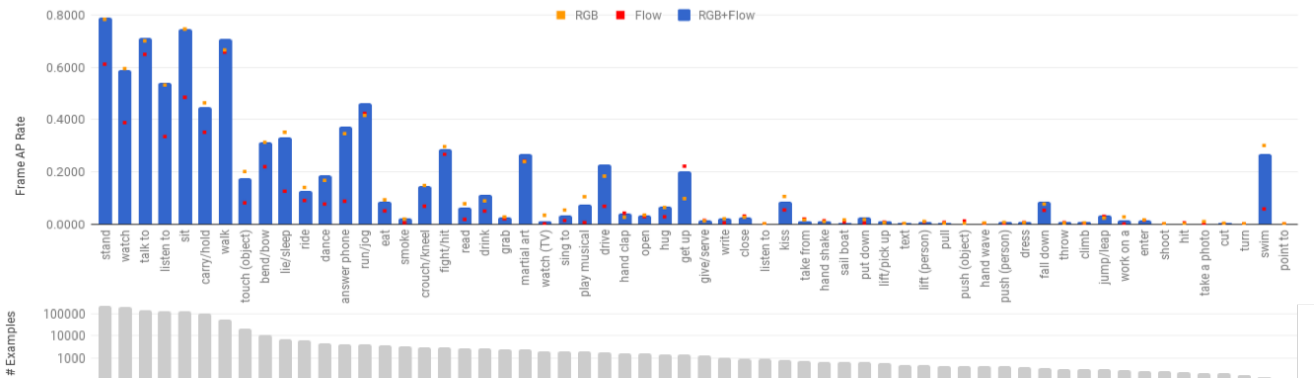

Figure 8. Top: We plot the performance of models for each action class, sorting by the number of training examples. Bottom: We plot the number of training examples per class. While more data is better, the outliers suggest that not all classes are of equal complexity. For example, one of the smallest classes "swim" has one of the highest performances because the associated scenes make it relatively easy.

Model	Temp.+ Mode	JHMDB	UCF101-24	AVA
2D	1 RGB + 5 Flow	52.1%	60.1%	14.2%
3D	5 RGB + 5 Flow	67.9%	76.1%	13.6%
3D	10 RGB + 10 Flow	73.4%	78.0%	14.2%
3D	20 RGB + 20 Flow	76.4%	78.3%	14.8%
3D	40 RGB + 40 Flow	76.7%	76.0%	**15.8%**
3D	50 RGB + 50 Flow	-	73.2%	15.7%
3D	20 RGB	73.2%	77.0%	14.6%
3D	20 Flow	67.0%	71.3%	10.1%

Table 4. Frame-mAP @ IoU 0.5 for action detection on JHMDB (split1), UCF101 (split1) and AVA. Note that JHMDB has up to 40 frames per clip. For UCF101-24, we randomly sample 20,000 frame subset for evaluation. Although our model obtains state-of-the-art performance on JHMDB and UCF101-24, the fine-grained nature of AVA makes it a challenge.

	JHMDB	UCF101-24	AVA
Action detection	76.7%	76.3%	15.8%
Actor detection	92.8%	84.8%	75.3%

Table 5. Frame-mAP @ IoU 0.5 for action detection and actor detection performance on JHMDB (split1), UCF101-24 (split1) and AVA benchmarks. Since human annotators are consistent, our results suggest there is significant headroom to improve on recongizing atomic visual actions.

challenging on AVA than on JHMDB, the gap between localization and end-to-end detection performance is nearly 60% on AVA, while less than 15% on JHMDB and UCF101. This suggests that the main difficulty of AVA lies in action classification rather than localization. Figure 9 shows examples of high-scoring false alarms, suggesting that the difficulty in recognition lies in the fine-grained details.

Which categories are challenging? How important is number of training examples? Figure 8 breaks down performance by categories and the number of training examples. While more data generally yields better performance, the outliers reveals that not all categories are of equal complexity. Categories correlated with scenes and objects (such as swimming) or categories with low diversity (such as fall down) obtain high performance despite having fewer training examples. In contrast, categories with lots of data,

Figure 9. Red boxes show high-scoring false alarms for smoking. The model often struggles to discriminate fine-grained details.

such as touching and smoking, obtain relatively low performance possibly because they have large visual variations or require fine grained discrimination, motivating work on person-object interaction [7, 12]. We hypothesize that the gains on recognizing atomic actions will need not only large datasets, such as AVA, but also rich models of motion and interactions.

7. Conclusion

This paper introduces the AVA dataset with spatio-temporal annotations of atomic actions at 1 Hz over diverse 15-min. movie segments. In addition we propose a method that outperforms the current state of the art on standard benchmarks to serve as a baseline. This method highlights the difficulty of the AVA dataset as its performance is significantly lower than on UCF101 or JHMDB, underscoring the need for developing new action recognition approaches.

Future work includes modeling more complex activities based on our atomic actions. Our present day visual classification technology may enable us to classify events such as "eating in a restaurant" at the coarse scene/video level, but models based on AVA's fine spatio-temporal granularity facilitate understanding at the level of an individual agents actions. These are essential steps towards imbuing computers with "social visual intelligence" – understanding what humans are doing, what they might do next, and what they are trying to achieve.

Acknowledgement We thank Abhinav Gupta, Abhinav Shrivastava, Andrew Gallagher, Irfan Essa, and Vicky Kalogeiton for discussion and comments about this work.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *arXiv:1609.08675*, 2016. 2

[2] D. Arijon. *Grammar of the film language*. Silman-James Press, 1991. 2

[3] R. Barker and H. Wright. *Midwest and its children: The psychological ecology of an American town*. Row, Peterson and Company, 1954. 2

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005. 2

[5] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *CVPR*, 2017. 2, 3, 6

[7] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 3, 8

[8] K.-W. Church and P. Hanks. Word association norms, mutual information, and lexicoraphy. *Computational Linguistics*, 16(1), 1990. 5

[9] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge: A retrospective. *IJCV*, 2015. 3, 7

[10] Geena Davis Institute on Gender in Media. The Reel Truth: Women Aren't Seen or Heard. https://seejane.org/research-informs-empowers/data/, 2016. 3

[11] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, 2015. 3

[12] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 2, 8

[13] S. Gupta and J. Malik. Visual semantic role labeling. *CoRR*, abs/1505.04474, 2015. 3

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[15] G. V. Horn and P. Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv:1709.01450*, 2017. 2

[16] R. Hou, C. Chen, and M. Shah. Tube convolutional neural network (T-CNN) for action detection in videos. In *ICCV*, 2017. 2, 3, 6, 7

[17] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017. 6

[18] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". *CVIU*, 2017. 2

[19] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 6

[20] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. Black. Towards understanding action recognition. In *ICCV*, 2013. 2, 3, 7

[21] 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. 2

[22] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics human action video dataset. *arXiv:1705.06950*, 2017. 2, 6

[23] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005. 3

[24] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *ICCV*, 2011. 2

[25] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955. 4

[26] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2

[27] P. Mettes, J. van Gemert, and C. Snoek. Spot On: Action localization from pointly-supervised proposals. In *ECCV*, 2016. 3

[28] M. Monfort, B. Zhou, S. A. Bargal, T. Yan, A. Andonian, K. Ramakrishnan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick, et al. Moments in time dataset: one million videos for event understanding. 2

[29] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. Smeaton, and G. Quénot. TRECVID 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics, 2014. 2

[30] X. Peng and C. Schmid. Multi-region two-stream R-CNN for action detection. In *ECCV*, 2016. 3, 6, 7

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 3, 4, 6

[32] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 2, 3

[33] S. Saha, G.Sing, and F. Cuzzolin. AMTnet: Action-micro-tube regression by end-to-end trainable deep architecture. In *ICCV*, 2017. 3

[34] S. Saha, G. Singh, M. Sapienza, P. Torr, and F. Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. In *BMVC*, 2016. 3

[35] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004. 2

[36] G. Sigurdsson, O. Russakovsky, A. Farhadi, I. Laptev, and A. Gupta. Much ado about time: Exhaustive annotation of temporal data. In *Conference on Human Computation and Crowdsourcing*, 2016. 4

[37] G. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2, 6

[38] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *ICCV*, 2017. 3, 7

[39] K. Soomro, A. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. Technical Report CRCV-TR-12-01, University of Central Florida, 2012. 2, 3

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 6

[41] V.Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017. 2, 3, 6, 7

[42] L. Wang, Y. Qiao, X. Tang, and L. Van Gool. Actionness estimation using hybrid fully convolutional networks. In *CVPR*, 2016. 7

[43] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, 2015. 3

[44] P. Weinzaepfel, X. Martin, and C. Schmid. Towards weakly-supervised action localization. *arXiv:1605.05197*, 2016. 3

[45] L. Wu, C. Shen, and A. van den Hengel. PersonNet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016. 4

[46] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *IJCV*, 2017. 2

[47] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, 2009. 3

[48] H. Zhao, Z. Yan, H. Wang, L. Torresani, and A. Torralba. SLAC: A sparsely labeled dataset for action classification and localization. *arXiv preprint arXiv:1712.09374*, 2017. 2

[49] M. Zolfaghari, G. Oliveira, N. Sedaghat, and T. Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *ICCV*, 2017. 3