HAA300: A New Benchmark for Human-Centric Atomic Action Recognition

Anonymous CVPR submission

Paper ID

Abstract

We propose HAA300: human-centric atomic action video dataset for fine-grained action recognition. Unlike existing and emerging atomic action datasets [15, 5, 13, 34] emphasizing data size and videos captured "in the wild," HAA300 advocates to human-centric actions with an average of 77% detectable joints for the relevant human poses. Our dataset contains 300 human action classes with 6K action labels in total. Each clip (taken from Youtube) lasts no more than 3 seconds. The collected human actions cover a broad range of classes including human-only actions such as backflip, human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands. We describe the dataset statistics, collection methodology, and compare quantitative performance for baseline neural network architectures trained and tested on human action classification task using this dataset. We demonstrate fine-grained action recognition by evaluating the AVA dataset with models pretrained with HAA300, and transfer learning results on UCF101 and HMDB51 after fine-tuning our pretrained model.

1. Introduction

Large-scale image datasets such as ImageNet [7] and COCO [25] have greatly contributed to the recent breakthrough in image understanding especially detection and recognition. However, we have not witnessed the same level of impact on video understanding contributed by large-scale datasets in particular action recognition. Observe the *coarse* annotation provided by commonly-used action recognition datasets such as [28, 21, 50], where the same action label was assigned to a given complex video action sequence (e.g. "Play Football," "Play Rugby") typically lasting 10 seconds and thus 300 frames, thus introducing a lot of ambiguities during training as two or more action categories may contain the same *atomic action* (e.g., "Run" is one of the atomic actions for both "Play Football" and "Play Rugby"). In this paper we are interested in human-centric atomic actions (HAAs):

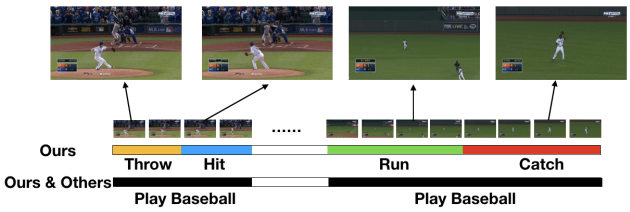

Figure 1. Recognizing "Play baseball" with its atomic actions using our approach vs others.

Human-centric Atomic Action (HAA) Unlike [15], our human-centric atomic action (HAA) videos were curated with an average of 77% detectable joints and 18% human coverage, while similar to [15] they have clear visual signatures, typically independent of the interacted objects. Such atomic actions usually last no longer than 3 seconds, which are generally enough for capturing a fine-scale temporal movement.

To address the issue of fine-grained action recognition, a number of emerging *atomic action* datasets [15, 5, 13, 34] have been recently introduced. Google's AVA actions dataset [15] provides dense annotations of 80 atomic visual actions in 430 fifteen-minute video clips, where actions are localized in space and time. Later, AVA spoken activity dataset [35] contains temporally labeled face tracks in videos, where each face instance is labeled as speaking or not, and whether the speech is audible. The something-something dataset [13] contains clips of humans performing pre-defined basic actions with everyday objects.

While the general direction is commendable in decomposing complex actions into a sequence of atomic actions, observed that the videos collected in the first versions of these datasets are almost "in the wild" and the reported performance was very low [15]. Current state-of-the-art performance on AVA is only 34% mAP. We believe AVA's limited advance in video recognition with poor model performance is due to its action categorization in the presence of a large amount of noise, e.g., class "open" contains both actions such as "open a door" and "open a carpet" with totally different visuals, which poses a deep neural network an almost insurmountable challenge to properly learn the pertinent atomic action.

In this paper, we propose to collect curated action videos where the relevant human poses are detectable (2D pose skeletons as structure), and carefully construct our Human-centric Atomic Action datasets. HAA300 dataset provides fine-grained human-centric atomic visual action clips with clean temporal annotations. An example of the collected atomic actions is shown in Figure 1. The clips are class-balanced and contain clear visual signals with little occlusion. This clean, class-balanced dataset provides an effective alternative to advance research in atomic visual actions recognition and thus video understanding.

2. Related Works

2.1. Representative Action Recognition Dataset

Table 1 compares existing action recognition datasets. Representative action recognition datasets, such as KTH [36], HMDB51 [23], UCF101 [39], Weizmann [1], Hollywood-2 [26] and Kinetics [21], consist of short clips, which are manually trimmed to capture a single action. These datasets are ideally suited for training fully-supervised, whole-clip, forced-choice video classifiers. A few datasets used in action recognition research, such as CMU [22], MSR Actions [49], UCF Sports [33] and JHMDB [17], provide spatio-temporal annotations in each frame for short videos, but they only contain few actions. Aside from the subcategory of shortening the video length, recent extensions such as UCF101 [39], DALY [47] and Hollywood2Tubes [27] evaluate spatio-temporal localization in untrimmed videos, resulting in a performance drop due to the more difficult nature of the task. One common issue on these aforementioned datasets is their actions are restricted to a limited number of composite actions (e.g., tennis), thus are unsuitable for fine-grained action analysis and recognition. Google AVA [15] is the first large-scale atomic action dataset which provides 80 atomic action labels collected from movie clips each lasts for no longer than 3 seconds. However, for the action detection task in the AVA dataset, the state-of-the-art model only achieves mAP of 34% [11]. Other specialized datasets such as Moment in Times [28], HACS [50], Something-Something [13], and Charades-Ego [37] provide classes for atomic actions but none of them are HAA. Our dataset differs from all the aforementioned datasets as we provide over 300 fine-grained atomic human action classes with videos in higher quality (higher resolution), which is carefully constructed to have a better balanced class distribution, a finer granularity in video class selection, and more comprehensive action labeling.

2.2. Action Recognition Architectures

Current action recognition architectures can be categorized into two major approaches, using either 2D image-based or 3D video-based kernels for their convolutional and layer operators. [9, 16]. While 3D video-based methods achieve state-of-the-art performance on most of the action recognition tasks, the model size and total amount of parameters are considerably larger than 2D image-based methods.

Dataset	Videos	Actions	Atomic
KTH	600	6	
UCF Sports	182	10	
HMDB51	6,766	51	
UCF101	13,320	101	
Weizmann	81	10	
Hollywood-2	1,707	12	
Daily	36,000	10	
YouTube-8M	8,264,650	4,800	
Kinetics-400	306,245	400	
HACS	1,550,000	200	✓
Moment in Times	1,000,000	339	✓
AVA	57,600	80	✓
HAA300	**6,000**	**300**	✓

Table 1. Summary of representative action recognition datasets. HAA300 is human-labeled and human-centric (18% human coverage and 77% detectable joints) unlike other atomic action datasets.

2D Image-Based Approaches Temporally recurrent layers such as LSTMs and feature aggregation over time are two popular techniques for 2D image-based methods for propagating temporal information across frames. Without compromising the temporal structure and exploiting image classification networks [20, 24, 29, 30], ConvNets with LSTMs are considered as top performers among current models in this category. However, LSTMs suffer from unsatisfactory performance on feature extraction from low-level motion and expensive training cost, as backpropagation requires unrolling the network across multiple frames through time. Given this bottleneck, researchers have turned to the 3D video-based approaches for more complex action analysis tasks.

3D Video-Based Approaches By incorporating spatio-temporal filters to standard convolutional networks, 3D ConvNets are regarded as a natural approach to video modeling. However, a considerably large number of parameters needs to be optimized [18, 43, 44, 46, 8, 19, 40]. The more practical 3D video-based approach [38] is a two-stream network with both RGB frames and pre-computed optical flow as input to a ConvNet pretrained using ImageNet. Later, the two-stream inflated 3D ConvNets (I3D) [3] was proposed which combines 3D ConvNets and two-stream networks. This approach achieves a very competitive performance on existing benchmarks, and thus most of the state-of-the-art methods for action recognition have adopted this architecture [19]. While one line of research effort has been made to further improve the precision for the 3D video-

based approach [31, 6], the other focuses on finding the trade-off between precision and speed for this architecture, involving more effectively exploiting the spatial-channels and temporal-channels correlation information throughout network layers [45, 48, 51, 52, 42, 32].

2.3. Atomic Action Recognition

While there exist many action recognition datasets, most of them consist of composite actions (e.g. long-jump, high-jump, play-tennis, play-football) thus missing the possibility of utilizing the inherently hierarchical nature of a given action. In [10] the authors highlighted the hierarchical nature of human activities, where at the finest level, the actions consist of atomic body movements or object manipulation, while at the coarser level the most natural descriptions are in terms of intentionally and goal-directed behaviors. To model finer-level events, the AVA dataset [15] was introduced to provide person-centric spatio-temporal annotations on atomic actions, which includes 1.58M annotations in 80 atomic visual action categories. Although Google AVA is a large-scale dataset, in practice the dataset is too noisy for training purpose which is evidenced by the large number of submissions to the AVA challenge requiring pre-training on ImageNet or Kinetics-400 [12] to improve their model performance, while many entries still suffer from class imbalance and non-apparent actions due to low-resolution videos in the AVA dataset. On the other hand HAA300 provides a clean and class-balanced dataset on fine-grained actions [4, 14].

3. HAA300

3.1. Data Collection

The annotation of HAA300 consists of two stages: vocabulary collection and video clip selection. Unlike Google AVA's bottom-up approach which generates action classes based on the selected videos, we aim at building a clean and fine-grained dataset for atomic action recognition, thus the videos are collected based on pre-defined atomic actions following a top-down approach.

Vocabulary Collection To make the dataset as clean as possible and useful for recognizing fine-grained atomic actions, we narrowed down the scope of our super classes into sports, instrument-playing, and daily routines. Next, following the definition of atomic actions, our dataset has a hierarchical structure: a composite action as root and different atomic actions decomposed from its composite category as leaves. For example, *playing tennis* is a composite action containing leaves such as *serving, swinging forehand, swinging backhand, volleying forehand and volleying backhand*. Such design poses a challenge on action recognition as we introduce both generality and diversity among the actions. Our dataset has atomic actions that vary within its root action while overlapping across different general actions (e.g., forehand and serve are significantly different under their root action of playing tennis, but forehand itself appears similarly across general actions of playing tennis and playing table tennis). We chose our classes to be human-centric and pose-apparent because these choices not only result in relatively less occlusion but also enable an alternative approach to recognizing actions with human-pose-based information. Consequently, we ended up with 300 atomic action classes, where 178 are sports, 43 are instrument playing and the rest are daily routines. The appendix in Figure 9 shows some samples classes in HAA300.

Video Clip Selection To ensure our dataset is clean and class-balanced, all the video clips are collected from YouTube with the majority having a resolution of at least 720p. Each class of atomic action contains at least 16 training clips. We manually selected the clips with apparent human-centric actions for each class from the top-50 Youtube videos returned by searching using the class name. Clips are properly trimmed to cover the desired actions. In general, each class contains clips from at least 3 different sources so that the action recognition task cannot be trivially reduced to identifying the corresponding background.

Training/Validation/Test Sets They are split at the class level, and since the clips in different classes are mutually exclusive, all clips appear only in one split. The 6000 clips are split roughly as 16:1:3, resulting in segments of 4800 training, 300 validation, and 900 test clips.

3.2. Dataset Analysis

HAA300 includes 300 atomic action classes where each class contains around 20 clips. Its human-centric and fine-grained nature are the two key features differentiating HAA from other existing datasets.

3.2.1 Statistics of HAA300

Table 2 summarizes the HAA300 statistics. The dataset contains around 6000 clips with an average length of 1.49 second. Figure 2 shows the distribution of clip lengths ($<$ 1s, 1 – 2s, or 2 – 3s) and the average clip length (blue line) for each class. In addition to action category labels, each clip was annotated with several meta-information tags to allow a more precise and thorough evaluation. The meta-information contains the following fields: the number of persons (single or multiple), the presence of occlusion (human-verified), and the consistency of camera angle. Figure 3 shows the distribution of the meta tags for the dataset.

action	clips	mean clip length	total duration
300	6000	1.46s	8760s

Table 2. Summary of HAA300

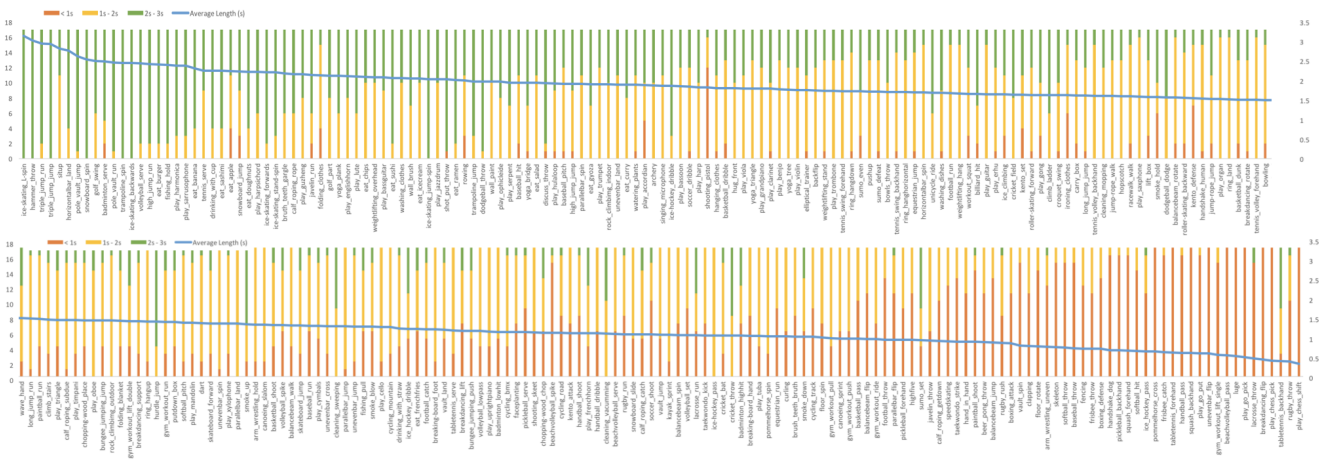

Figure 2. Distribution of clip lengths (<1s, 1 – 2s, or 2 – 3s) and average clip length (blue line)

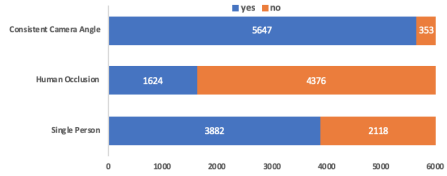

Figure 3. Meta-information for the whole dataset .

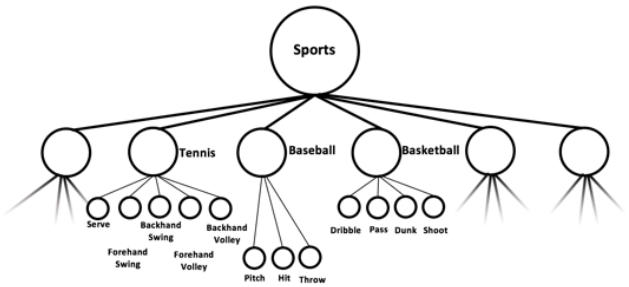

Figure 4. Tree-diagram of classes: sports branch as an example. The full hierarchical structure is shown in Figure 10 in the Appendix.

3.2.2 Characteristics of HAA300

Fine-grained Each class under the HAA300 hierarchy is mutually exclusive, with only the leaf nodes counted as the 300 classes. In general, the first level of the hierarchy is categorized into "sport", "play instrument", and "daily routine". The second level consists of, for example, "tennis", "badminton" and "kendo" categories, where each sport type will then be further broken into the corresponding atomic actions. For example, tennis is composed of *serving*, *swing forehand*, *swing backhand*, *volley forehand* and *volley backhand* in the third level. Figure 4 illustrates part of the class heirarchical structure.

Human-centric Every clip in HAA300 is human-centric with detectable human poses and thus conducive to fine-grained action recognition. The average of the human coverage in each clip across the whole training set is 18%, where single-person clips tend to have a higher coverage. We also analyze the percentage of detectable joints (each skeleton has a maximum 18 joints) in each clip in the training set and compute the average number of detectable joints for different classes with a keypoint detection model [2]. The number of detectable joints and human coverage do not correlate, while generally sport classes have a larger number of detectable joints. Figure 5 shows the distribution of average human coverage and number of detectable joints, with an average of 77% detectable joints in our dataset.

4. Experiments

4.1. Baseline Benchmarks

We have evaluated the baseline performance on our HAA300 with three different architectures. We adopted 2D-based CNNs and 3D-based CNNs including the I3D network and the SlowFast Network. Our experiments were implemented with PyTorch.

4.1.1 Implementation Details

We adopted the pre-trained ResNet-101 [16] architecture for the 2D models. Figure 6 shows the model architecture. One frame in each clip is randomly selected as the input to the 2D model. In the 3D models, we extracted RGB frames, optical flows, and human-pose estimation for each frame in the clips as inputs. We randomly selected 64 consecutive frames from the clip. The frames were resized with the smallest dimension of 256 pixels with bilinear interpolation while preserving the aspect ratio. Pixel values were

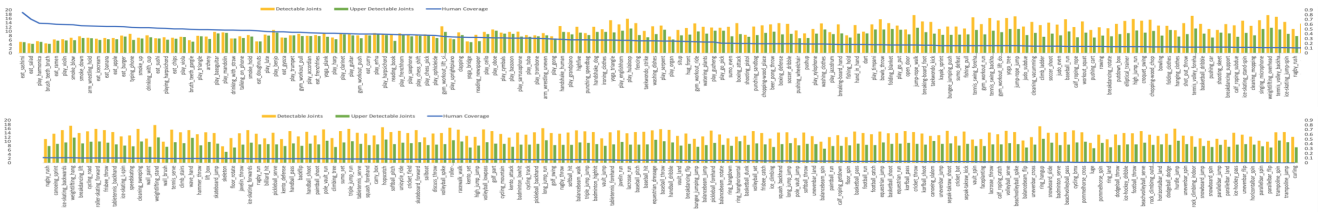

Figure 5. Average human coverage and number of detectable joints of different classes. Zoom in for details.

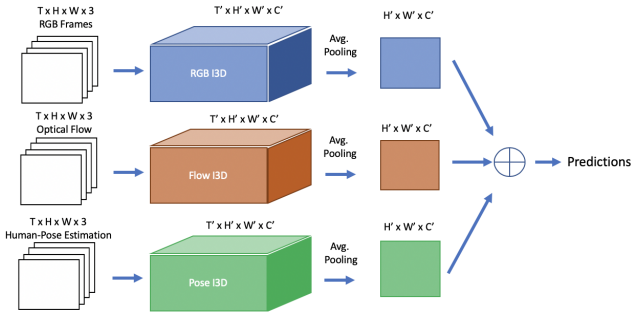

Figure 6. Baseline Model

rescaled to $[-1, 1]$. During training, we randomly selected a 224×224 image crop, while selecting the center 224×224 image crop from the video during testing. For the optical-flow stream, we converted the videos to grayscale and applied the PWC-Net [41] to extract the optical flows. The networks were trained with Softmax loss over 300 classes. Table 3 tabulates the classification results.

4.1.2 Results

	Top-1	Top-5	# of parameters
2D-CNN (RGB)	25.1	41.8	42M
I3D (RGB)	28.1	47.1	12M
I3D (Flow)	32.4	54.2	12M
I3D (Two-Stream)	36.8	59.5	12M
SlowFast (RGB)	29.2	45.5	62M

Table 3. Classification results on HAA300.

There are several noteworthy observations. First, the 3D-based model outperforms the 2D-based models. This suggests that temporal information is indeed a crucial factor for recognizing actions, since multiple actions may share similar human-poses at particular frames or contain similar background. Second, the contribution from the flow stream is slightly higher than that from the RGB stream. We believe this results from our vocabulary collection methods, where multiple action classes are collected from similar scenes/backgrounds containing similar RGB information (See Figure 7). Optical flow, which captures motion features, can be helpful for discriminating actions among the same composite action.

(a) Balance-beam: Jump (b) Balance-beam: Rotate

Figure 7. Classes with similar backgrounds.

4.2. Human Pose Attention

With our human-centric dataset where human poses are detectable to a high extent, we evaluate the effectiveness of human pose estimations as an attention mechanism on HAA300.

4.2.1 Implementation Details

We extract human poses for each frame with OpenPose [2]. Outputs of OpenPose include a confidence map for body part detection and a map of Part Affinity Fields (PAFs), which is a set of 2D vector fields that encode the location and orientation of limbs over the image domain parts. We stack the confidence map together with the x and y components of the PAFs to form a three-dimensional representation map for each frame. We train the I3D model with the stacked feature maps, following the spatial and temporal cropping settings in our baseline model. Table 4 tabulates the classification results.

4.2.2 Results

Compared to the RGB-Stream I3D, the two-stream model with an additional human-pose input stream has improved the baseline performance for 10.5%. This suggests that human-pose information can be helpful for exploiting the human-centric nature of HAA300.

4.3. Fine-grained Action Labels

With the two-level action class label hierarchy of HAA300, we evaluate the effects of fine-grained labels to the performance of the coarser-grained actions.

	Top-1	Top-5	# of parameters
I3D (RGB)	28.1	47.1	12M
I3D (Human-Pose)	34.8	57.1	12M
I3D (Two-Stream)	39.6	63.2	12M

Table 4. Classification Results on HAA300

4.3.1 Fine-grained: HAA300

HAA300 contains 300 fine-grained action classes, each of which belongs to one of the 129 composite action classes. We compare the performance of models trained with and without an additional classification loss on the fine-grained action classes. We add an additional output stream to predict the fine-grained classes and sum up the classification loss on both the composite action classes and fine-grained action classes. We compare the performance on recognizing the 129 composite actions. We follow same settings in the baseline models for training. Table 5 tabulates the classification results.

4.3.2 Results

Method	Results
Composite classes	44.8
Composite + Fine-grained classes	**48.5**

Table 5. Results on two-level action class classification on HAA300

4.3.3 Fine-grained: AVA

We evaluate the effects of fine-grained action labels on the AVA dataset. We select a subset of the AVA dataset, containing the labels "Play musical instrument". We then classify each of the labels into two levels of finer-grained actions. The first level groups actions with similar visual signals, including "string instruments", "guitar instruments", "brass instruments", "woodwind instruments", "keyboard instruments", and "percussion instruments". The second level contains a further classification where each class represents playing a specific musical instrument, such as "play violin". This results in 6 coarser-level actions and 16 fine-grained actions. Table 6 shows the class structure. We follow the training settings in the previous section. Table 7 tabulates the classification results.

4.3.4 Results

Compared to models trained only with the composite labels, the model trained with an additional fine-grained classification loss has improved the performance of recognizing composite actions. This shows that such fine-grained labels enabled the model to learn better features in order to differentiate the subtle details that distinguishes each action from others.

Composite Action	Atomic Action
string instruments	play cello, play violin
guitar instruments	play guitar, play mandolin, play guitar, play guzheng
brass instruments	play trumpet, play trombone, play tuba, play frenchhorn, play saxophone
woodwind instruments	play clarinet, play harmonica
keyboard instruments	play piano, play accordion
percussion instruments	play drum

Table 6. Subset of composite actions in AVA

Method	Results
Composite classes	71.2
Composite + Fine-grained classes	**74.4**

Table 7. Results on two-level action class classification on AVA

4.4. Decomposition of Composite Actions: UCF101

We apply HAA300 to decompose the subset of composite actions in UCF 101 into different atomic actions. This experiment shows the capability of HAA300 as a fine-grained model can decompose complex composite actions which have different visual expressions in one clip into different atomic actions.

4.4.1 Implementation Details

We begin by selecting a subset of UCF101 that the subset will only contain composite action classes (eg. frisbee catch, baseball pitch, and volleyball spiking are several counterexamples). Table 8 shows the full subset of UCF101 we picked for this experiment. We select a subset of HAA300 that contains the corresponding actions in the UCF101 subset. We follow the training settings in the previous section. We evaluate the performance by feeding different segments of clips the chosen subset from UCF101.

4.4.2 Results

Among all the classes in the subset, HAA300 can successfully decompose classes Floor Gymnastic, Long Jump, High Jump, Parallel Bars, Pole Vault, Sumo Wrestling, and Uneven Bars into their corresponding atomic actions, while

Composite Action	Atomic Action
Balance Beam	spin, walk
Floor Gymnastic	spin, rotate
Long Jump	run, jump
High Jump	run, jump
Jump Rope	jump, walk
Sumo Wrestling	even, defeat
Parallel Bars	spin, jump, land
Uneven Bars	spin, jump, land
Pole Vault	run, jump
Trampoline Jump	jump, spin

Table 8. Subset of composite actions in UCF1010

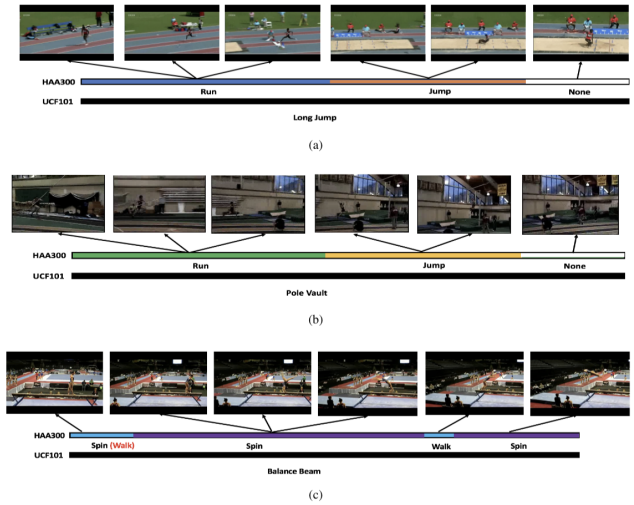

Figure 8. (a) and (b) are examples of successful classes that can be decomposed by HAA300. (c) is an example of failed class that HAA300 wrongly classifies atomic action walk to spin.

HAA300 fails to decompose classes Jump Rope, Balance Beam, and Trampoline Jump. Figure 8 shows the result of several successful and failed examples.There are 2 observations we identify for explaining the experiment results. Firstly, for the successful classes, we found out that their composite action can all be broke down into atomic actions that are in absolute sequence and are significantly different in visual expression. For example, composite action Floor Gymnastic is always composed by two atomic actions spinning then rotating. Same cases for High Jump, Long Jump and Pole Vault, all of them are always composed by two atomic actions running then jumping. While for the failed classes Jump Rope, Balance Beam, and Trampoline Jump, they are not composed by several atomic actions in absolute order. For examples, atomic actions walking and spinning can appear alternatively during the whole composite action Balance Beam. This difference between successful classes and failed classes indicates that a crucial component for HAA300 to decompose a composite action is the temporal information. Secondly, we found out that the difference of visual expression between each atomic actions for composite actions in failed class is not clear and can sometime overlap with each other. For examples, atomic actions walking sometime will be included in the beginning of spinning for composite action Balance Beam, thus make the model be confused with the boundary of different atomic actions.

4.5. AVA Action Detection

We evaluate the generalization performance of models trained with HAA300 with a subset of the AVA dataset. We show that the fine-grained nature of HAA300 can improve performance on composite action recognition.

4.5.1 Implementation Details

We begin by selecting a subset of the AVA dataset, containing the frames with labels "Run", "Play Instruments", and "Eat" from both the training set and the validation set. We selected these action categories since these actions are typically exclusive of each other, e.g. a person would normally not be "running" and "playing instrument" at the same time. To better assess the model's ability to recognizing human-centric actions, we removed frames from the subset that contains excessive occlusion. This results in a subset of 7659 training clips and 766 validation clips, which we dub as **miniAVA**. We then select a subset of actions from HAA300 under categories "Run", "Play Instruments", and "Eat", which contains 31 actions to produce **miniHAA**. We train a model with miniHAA following the implementation details in the baseline model. We compare the performance with a model trained from scratch with miniAVA and report the mean Average Precision (mAP) over the three classes. Table 9 tabulates the results.

4.5.2 Results

Compared to the model trained on miniAVA, the model trained on miniHAA has improved the mAP for 4.9%. The result seems even more surprising considering the fact that the size of miniHAA ($\sim$ 600) is much smaller than the training set in miniAVA ($\sim$ 7000). The suggests that a class-balanced dataset of human actions in HAA300 helps the model to learn the features in a more efficient manner even HAA300 is much smaller. Thus, HAA300 is easily extensible to cover more classes with fewer videos required while achieving better performance.

4.6. Transfer Learning Results

Models trained on HAA300 can be finetuned on other action datasets. By comparing finetuned models with models trained from scratch, we can assess the generalization performance of spatial-temporal features learned on our

	miniAVA	**miniHAA**
Play Instrument	36.8	34.7
Eat	45.9	55.6
Run	86.4	93.5
mAP	56.4	**61.3**

Table 9. Results on miniAVA

dataset. We evaluate the transfer learning results on 2 action classification benchmarks, UCF101 and HMDB51.

4.6.1 Implementation Details

We train and test on the I3D model with RGB inputs and optical flow inputs. We begin by evaluating the performance of models trained from scratch with UCF101 and HMDB51. We then evaluate the models pretrained with HAA300 followed by finetuning on the two datasets. We follow the spatial and temporal cropping settings in the baseline models. Table 10 tabulates the results.

4.6.2 Results

The results showed that models pretrained on HAA300 has an improved performance on the HMDB51 dataset, and a marginal gain difference on the performance of UCF101. The gains from pre-training with HAA300 on HMDB51 may be higher due to the nature of the dataset, where the average clip length (∼3 seconds) is closer to that of HAA300 compared to UCF101 (∼7 seconds).

Model	**Pretrain**	UCF101 Top-1	HMDB51 Top-1
I3D (RGB)	None	58.5	22.2
I3D (RGB)	HAA300	**70.9**	**39.1**
I3D (Flow)	None	85.2	50.7
I3D (Flow)	HAA300	**82.8**	**53.5**
I3D (Two-Stream)	None	85.5	53.1
I3D (Two-Stream)	HAA300	**85.6**	**58.6**

Table 10. Performance on the UCF-101 and HMDB-51 test sets (split 1 of both) starting with / without HAA300 pretrained weights.

5. Conclusion

This paper introduces HAA300, a new human action dataset with fine-grained atomic action labels and human-centric clip annotations. We have provided baseline benchmarks on this dataset and demonstrated the effectiveness of human-pose estimations on improving the performance of recognizing human actions. We have also shown how fine-grained action labels can help recognizing composite actions decompose composite actions into atomic atomic actions. We hope HAA300 will inspire new methods and architectures for modeling higher-complexity human actions.

6. Appendix

Figure 9 shows some sample classes in HAA300.

Figure 9. Sample classes in HAA300.

Figure 10 shows the full hierarchy of HAA300.

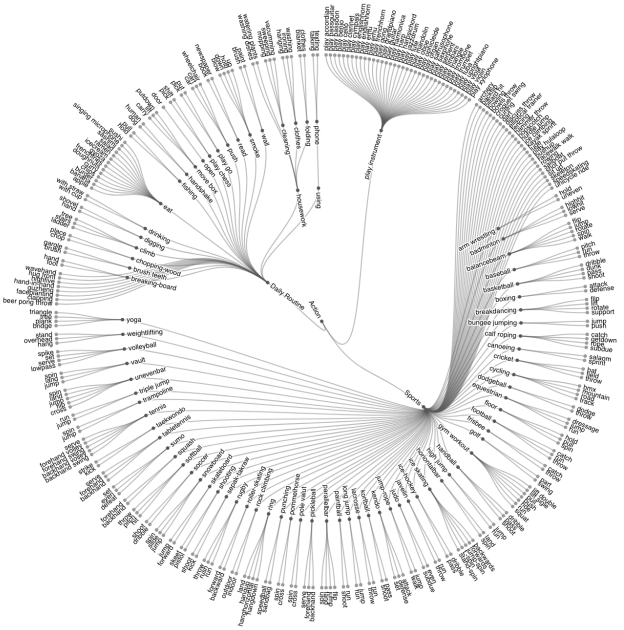

Figure 10. Full hierarchical structure of all classes in HAA300.

References

[1] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *2005 IEEE International Conference on Computer Vision (ICCV 2005)*, 2005. 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 4, 5

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017. 2

[4] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *2015 IEEE International Conference on Computer Vision (ICCV 2015)*, 2015. 3

[5] Sourish Chaudhuri, Joseph Roth, Daniel P. W. Ellis, Andrew C. Gallagher, Liat Kaver, Rebecca Marvin, Caroline Pantofaru, Nathan Reale, Loretta Guarino Reid, Kevin W. Wilson, and Zhonghua Xi. Ava-speech: A densely labeled dataset of speech activity in movies. In *INTERSPEECH*, 2018. 1

[6] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, 2019. 3

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[8] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Mohammad Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool. Spatio-temporal channel correlation networks for action classification. In *2018 European Conference on Computer Vision (ECCV 2018)*, 2018. 2

[9] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *2015IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 2015. 2

[10] H. Warren Dunham. Midwest and its children: The psychological ecology of an american town. roger g. barker , herbert f. wright. *American Journal of Sociology*, 62(2), 1956. 3

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018. 2

[12] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. A better baseline for AVA. *CoRR*, abs/1807.10066, 2018. 3

[13] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, 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. In *ICCV*, volume 1, page 3, 2017. 1, 2

[14] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE International Conference on Computer Vision (ICCV 2017)*, 2017. 3

[15] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 2018. 1, 2, 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, 2016. 2, 4

[17] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *2013 IEEE International Conference on Computer Vision (ICCV 2013)*, 2013. 2

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. In *2010 International Conference on Machine Learning (ICML 2010)*, 2010. 2

[19] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. STM: spatiotemporal and motion encoding for action recognition. *CoRR*, abs/1908.02486, 2019. 2

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014. 2

[21] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1, 2

[22] Yan Ke, Rahul Sukthankar, and Martial Hebert. Efficient visual event detection using volumetric features. In *2005 IEEE International Conference on Computer Vision (ICCV 2005)*, 2005. 2

[23] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *2011 IEEE International Conference on Computer Vision (ICCV 2011)*, 2011. 2

[24] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In

David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1

[26] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 2009. 2

[27] Pascal Mettes, Jan C. van Gemert, and Cees G. M. Snoek. Spot on: Action localization from pointly-supervised proposals. In *2016 European Conference on Computer Vision (ECCV 2016)*, 2016. 2

[28] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *CoRR*, abs/1801.03150, 2018. 1, 2

[29] Juan Carlos Niebles, Hongcheng Wang, and Fei-Fei Li. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3), 2008. 2

[30] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, 2014. 2

[31] Bowen Pan, Jiankai Sun, Wuwei Lin, Limin Wang, and Weiyao Lin. Cross-stream selective networks for action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (CVPR 2017)*, 2019. 3

[32] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *IEEE International Conference on Computer Vision (ICCV 2017)*, 2017. 3

[33] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008. 2

[34] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. Ava-activespeaker: An audio-visual dataset for active speaker detection, 2019. 1

[35] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru. Ava-activespeaker: An audio-visual dataset for active speaker detection. *arXiv:1901.01342*, 2019. 1

[36] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *2004 IEEE International Conference on Pattern Recognition (CVPR 2004)*, 2004. 2

[37] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016. 2

[38] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *2014 Neural Information Processing Systems (NIPS 2014)*, 2014. 2

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 2

[40] Jonathan C. Stroud, David A. Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3D: distilled 3d networks for video action recognition. *CoRR*, abs/1812.08249, 2018. 2

[41] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[42] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E. Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV 2015)*, 2015. 3

[43] Graham W. Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *2010 European Conference on Computer Vision (ECCV 2010)*, 2010. 2

[44] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV 2015)*, 2015. 2

[45] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 2018. 3

[46] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1510–1517, 2018. 2

[47] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Towards weakly-supervised action localization. *CoRR*, abs/1605.05197, 2016. 2

[48] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *2018 European Conference on Computer Vision (ECCV 2018)*, 2018. 3

[49] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative subvolume search for efficient action detection. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 2009. 2

[50] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. Hacs: Human action clips and segments dataset for recognition and temporal localization. *arXiv preprint arXiv:1712.09374*, 2019. 1, 2

[51] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Mict: Mixed 3d/2d convolutional tube for human action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 2018. 3

[52] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. ECO: efficient convolutional network for online video understanding. In *2018 European Conference on Computer Vision (ECCV 2018)*, 2018. 3