Cultivating HAA150: Human-Centric Atomic Action Dataset with Curated Videos

Anonymous ECCV submission

Paper ID 3436

Abstract. We contribute HAA150, a manually annotated human-centric atomic action dataset for action recognition on 150 classes with over 240k labeled frames. Unlike existing atomic action datasets some of which were collected "in the wild," HAA150 has been carefully curated to capture movement of human figures with less spatio-temporal noises to greatly enhance the training of deep neural networks. The advantages of HAA150 include: 1) human-centric actions with a high average of 75.1% detectable joints for the relevant human poses; 2) high sampling frequency which effectively ameliorates the adverse effect of temporal noises; 3) fine-grained atomic action classes. Our extensive experiments have validated that with its clean and fine-grained nature, HAA150 significantly improves action recognition tasks even only adopting baseline architecture under different settings. We detail the HAA150 dataset statistics and collection methodology, and compare quantitatively with existing action recognition datasets.

Keywords: Action Recognition; Atomic Actions; Human-Centric; Video Understanding

1 Introduction

Large-scale image datasets such as ImageNet [7] and COCO [23] 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 [19, 26, 47], where the same action label was assigned to a given complex video action sequence (e.g. "Play Soccer," "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 Soccer" and "Play Rugby").

Recently, atomic action datasets [5, 12, 13, 33] have been introduced in attempt to resolve the aforementioned issue. Google's AVA actions dataset [13] provides dense annotations of 80 atomic visual actions in 430 fifteen-minute

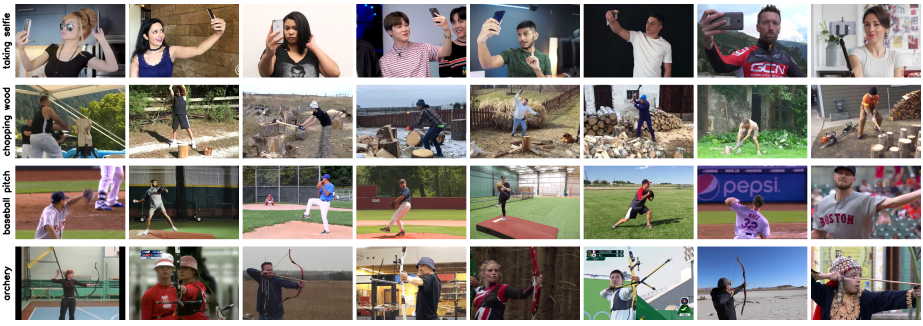

Fig. 1. Sample videos of HAA150 for corresponding action classes. Every video contain only one or two dominant human figures performing the labeled atomic action.

video clips, where actions are localized in space and time. Later, AVA spoken activity dataset [32] 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 [12] contains clips of humans performing pre-defined basic actions with everyday objects.

However, some of their actions are still coarse which can be further split into atomic classes that have far different motion gestures. e.g. AVA[13] and something-something[12] contain "Play Musical Instrument" and "Throw Something" as a class, respectively, where the former should be further divided into sub-classes such as "Play Piano" and "Play Guitar", and the latter into "Throw Frisbee" and "Pitch Baseball", etc, because each of these atomic actions have significantly different gestures. Encompassing different visual postures into a single atomic class poses a deep neural network almost insurmountable challenge to properly learn the pertinent atomic action, which probably explains the prevailing low performance employing even the most state-of-the-art architecture, SlowFast(mAP: 34.2%) [10], in AVA [13].

The other problem with action recognition video datasets is their training examples contain strong noises or actions irrelevant to the target action. Video datasets tend to have low sampling frequencies, allowing unrelated video frames to be easily included during the data collection stage. Kinetics400 dataset [19], with a sampling frequency of 1 Hz, exhibit a lot of action noises, e.g., beginning frames showing the audience before the main violin playing action, or a person running before kicking the ball. Field-of-view is another problem, with videos in existing action recognition datasets only exhibit part of a human interacting with an object, e.g. a hand [12], or multiple human figures with different actions present [13, 19, 47].

This paper introduces Human-centric Atomic Action dataset (**HAA150**) that has been constructed with carefully curated videos with an average of 75.1% detectable joints and 20.2% human coverage, where a dominant human figure is present to perform the labeled action. Videos have been annotated with fine-grained labels to avoid ambiguity, and with a sampling frequency of a single frame to avoid unrelated video frames being included in the annotation. The

clips are class-balanced and contain clear visual signals with little occlusion. As opposed to "in the wild" atomic action datasets, our "cultivated" clean, class-balanced dataset provides an effective alternative to advance research in atomic visual actions recognition and thus video understanding. An example of the collected atomic actions is shown in Figure 1.

2 Related Works

2.1 Action Recognition Dataset

Representative action recognition datasets, such as KTH [34], HMDB51 [21], UCF101 [37], Weizmann [2], Hollywood-2 [24] and Kinetics [19], 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 [20], MSR Actions [46], UCF Sports [31] and JHMDB [15], 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 [37], DALY [44] and Hollywood2Tubes [25] 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 that they are annotated with composite action classes (e.g. tennis), thus causing multiple action classes (e.g. backhand swing, forehand swing) to be annotated under a single class. Another issue is that they tend to have wide field-of-view, including multiple human figure (e.g. tennis player, referee, audience) with different actions in a single frame, thus are unsuitable for action analysis and recognition. Table 1 tabulates summary of representative action recognition datasets.

Table 1. Summary of representative action recognition datasets.

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	✓
HAA150	**3,638**	**150**	✓

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, 14]. While 3D video-based methods achieve state-of-the-art performance on most of the action recognition tasks, the model size and total number of parameters are considerably larger than 2D image-based methods.

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 [18, 22, 27, 28], 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 [8, 16, 17, 38, 40, 41, 43]. The more practical 3D video-based approach [36] 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) [4] 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 [17]. While one line of research effort has been made to further improve the precision for the 3D video-based approach [6, 29], 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 [30, 39, 42, 45, 48, 49].

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, tennis, football) thus missing the possibility of utilizing the inherently hierarchical nature of a given action. In [1] 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 [13] 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 [11] 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. Other specialized datasets such as Moments in Times [26], HACS [47], Something-Something [12], and Charades-Ego [35] provide classes for atomic actions but none of them are Human-centric Atomic Action, where some of the videos are only showing part of a human body (e.g. hand), or no human action at all. Atomic action datasets [13, 26] tend to be have atomicity under English linguistics, e.g. *open* is annotated on video clips with tulip opening, eye opening, a person opening a door, and a person opening a package, which are are fundamentally different atomic actions with only similarity in sharing the verb *open*.

Our dataset differs from all the aforementioned datasets as we provide 150 fine-grained atomic human action classes that are fundamentally different from each other, carefully constructed to have a better balanced class distribution, a finer granularity in video class selection, and more comprehensive action labeling.

3 HAA150

3.1 Data Collection

The annotation of HAA150 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, where future extension beyond the existing classes is feasible. 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, *balance beam* is a composite action containing leaves such as *rotate, spin, walk, jump and flip*. 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., jump and spin are significantly different under their root action of balance beam, but spin itself appears similarly across general actions of balance beam and floor (gymnatics)). We chose our classes to be human-centric and pose-apparent because these choices not only result in

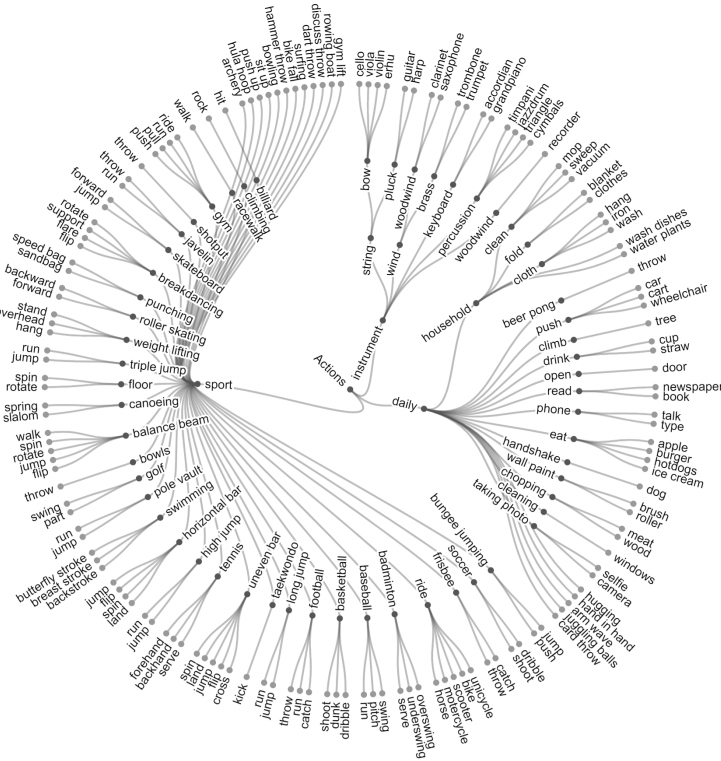

Table 2. HAA150 hierarchy.

Table 3. Summary of HAA150

action	clips	mean clip length	total duration	total number of frames
150	3,638	2.40s	8,719s	240K

relatively less occlusion but also enable an alternative approach to recognizing actions with human-pose-based information. Consequently, we ended up with 150 atomic action classes, where 94 are sports, 17 are instrument playing and the rest are daily routines. Figure 2 shows the classes of HAA150 and its hierarchy.

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 Youtube videos returned by searching using the class name. To increase diversity among the video clips, and avoid unwanted bias, all the clips were collected from different Youtube videos, with different environment setting so that the action recognition task cannot be trivially reduced to identifying

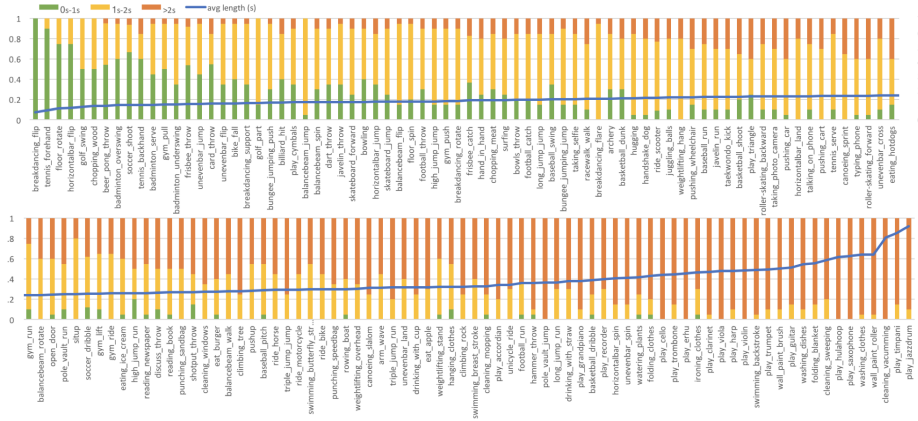

Fig. 2. Distribution of clip lengths (<1s, 1 – 2s, or 2s<) and average clip length (blue line)

the corresponding backgrounds. Clips are properly trimmed in a frame-perfect manner to cover the desired actions. Figure 9 shows examples of the selected video of each class.

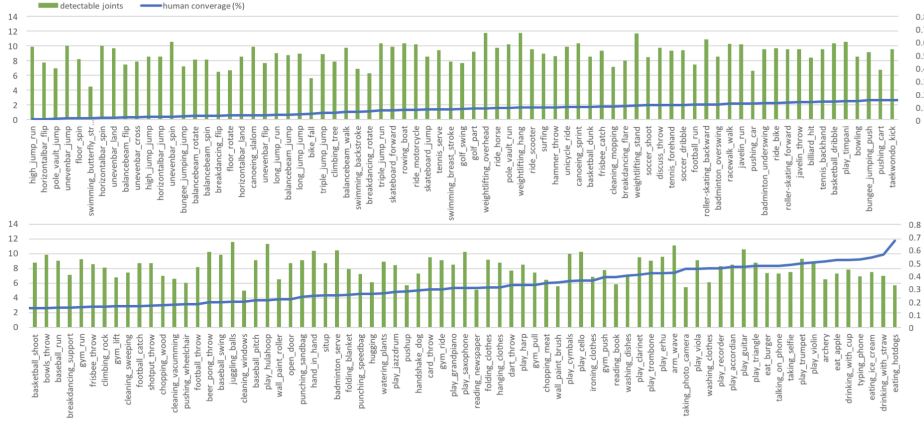

Fig. 3. Average human coverage and number of detectable joints of different classes. Zoom in for details.

Statistics HAA150 includes 150 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.

Table 3 summarizes the HAA150 statistics. The dataset contains around 3,638 clips with an average length of 2.40 seconds. Figure 2 shows the distribution

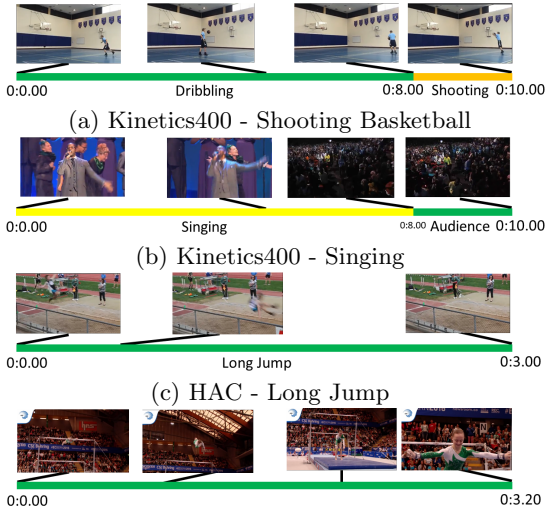

Fig. 4. Different types of action noise in action recognition datasets. **(a)**: Kinetics400 has sampling rate of 1 Hz and fixed video length of 10 seconds which cannot accurately annotate quick actions like "Shooting Basketball" where the irrelevant action of dribbling the ball is included in the clip. **(b)**: Camera cut can be seen, showing unrelated frames (audience) after the main action. **(c)**: HACS can only show part of "Long Jump", as HACS has a fixed length of 3 seconds, while "Long Jump" takes about 5 seconds to complete. **(d)**: Our HAA150 accurately annotates full motion of "Uneven Bars - Land" as it has high sample frequency and adjustable video length. From the exact frame the girl puts her hand off the bar, to the exact frame when she finishs her landing pose.

of clip lengths (0 – 1s, 1 – 2s, or 2s<) 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 dominant people in the video (2,584 videos contain 1 person; 354 videos contain 2 persons; 700 videos contain more than 2) and the consistency of camera angle. (2,775 videos were shot with moving camera; 863 videos with static camera).

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 3638 clips are split roughly as 16:1:3, resulting in segments of 2912 training, 178 validation, and 548 test clips.

3.2 Properties and Comparison

Clean Most video datasets [5, 13, 19, 33] have a low sampling frequency (1 Hz) for ease of collection and labeling. The original videos were annotated with an ac-

tion label at a temporal step size of 1 second which inevitably includes irrelevant action or noises in their dataset, thus making them unsuitable as benchmarks for action recognition. While HACS [47] has more precise temporal annotations, they do not distinguish two different actions under the same class when these actions are adjacent to each other in the source video clip. Table 4 tabulates the sampling rate of different video action datasets. Mentioned datasets also have fixed length of video clips, where temporal noises are inevitable for shorter or longer actions. Figure 4 shows examples of noises given from the constraints of other datasets. As HAA150 are curated to consist of perfect temporal annotation, we are totally free from any adverse effect due to these action noises.

Table 4. Sampling rate, clip length and scene distinction of video action datasets

Dataset	Sampling Rate	Clip Length	Single Scene
AVA	1 Hz	1 second	×
HACS	2 Hz	2 second	×
Kinetics400	1 Hz	10 second	×
Moments in Time	Random Cut	3 second	○
HAA150	Single Frame	adjustable	○

Fig. 5. The video clips in AVA, HACS, and Kinetics400 contain multiple human figure with different actions in the same frame. Something-Something focuses on the target object and barely shows any human body parts. In contrast, all video clips in HAA150 (in Figure 1) are carefully curated where each video shows either a single person, or the person-of-interest as the most dominant figure in a given frame.

Human-Centric One of the difficulties in action recognition is that the neural network tends to recognize by trivially comparing the background scene in the video, or detecting single key elements in a frame (e.g., a basketball to detect "Playing Basketball") rather than analyzing human gesture, thus causing the action recognition to have no better performance improvements over scene/object recognition. The other difficulty stems from the video action datasets where videos captured in wide field-of-view contain multiple people in a single frame [13, 19, 47], while videos captured using narrow field-of-view only exhibit very little

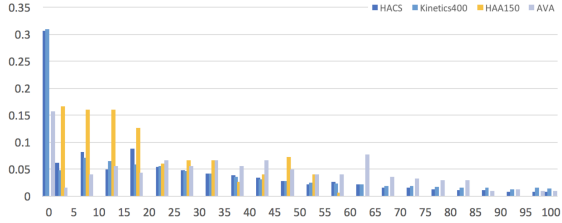

Fig. 6. Comparison of human coverage in different datasets. HAA150 has consistent human coverage from 5% to 50% and does not have any clips with 0% human coverage.

body part in interaction with an object [12, 26]. While in [13] attempts were made to overcome this issue through spatial annotation of each individual in a given frame, this introduces another problem of action localization and thus further complicating the difficult recognition task. Figure 5 illustrates example frames of different video action datasets.

HAA150 contributes a curated dataset where each human joint can be clearly detected, with consistent human-coverage over any given frame, thus allowing the model to benefit from learning human movements than just performing scene recognition. The average of human coverage in each clip across the whole training set is 20.2%. We also analyze the percentage of detectable joints (each skeleton has a maximum of 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 [3]. Figure 3 shows the distribution of average human coverage and number of detectable joints, with an average of 75.1% detectable joints in our dataset. Table 5 compares both detectable joints and human coverage over different video action datasets. Figure 6 shows that HAA150 contains consistent human coverage within 5% to 50%, with no clips at 0%. This is comparable to other datasets that peak at 0%, meaning they contain many video clips where human are not detectable at all. Other datasets vary greatly in human coverage, some go all the way to 100%.

Table 5. Detectable joints and body coverage percentage of video action datasets

Dataset	Detectable Joints	Human Coverage
AVA	61%	36.4%
Kinetics400	54.1%	25.7%
HACS	57.5%	22.7%
HAA150	**75.1%**	**20.2%**

Atomic A composite action is decomposed into its atomic action components for better recognition, e.g., action "Long Beam" contains "Jump", "Spin", and "Walk". While existing atomic action datasets [5, 13, 26] deconstruct composite actions to more basic atomic action elements, their atomicity is limited by

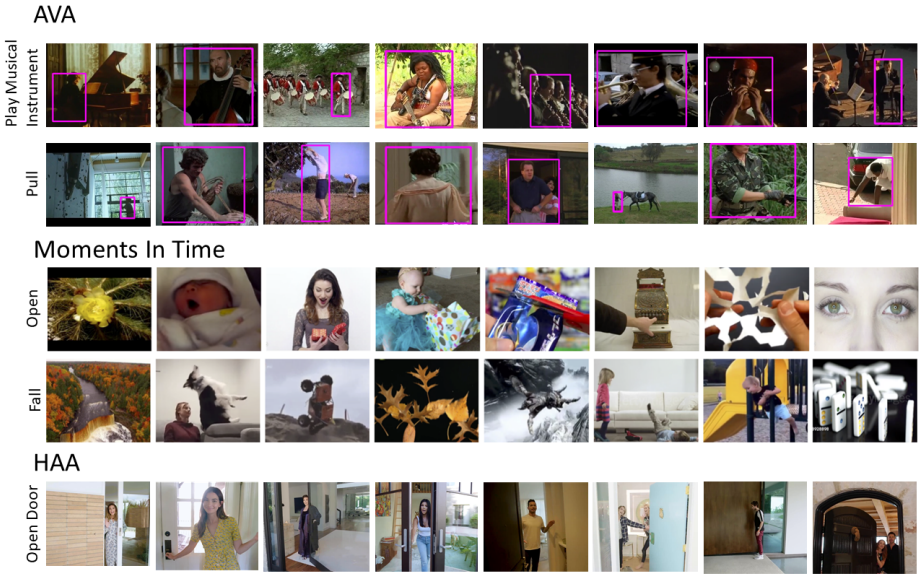

Fig. 7. Cases where atomic action datasets label different actions under a single English action verb. HAA150 (Bottom) has fine-grained classes where action ambiguities are eliminated as much as possible.

English linguistics, only decomposing to action verbs (e.g., walk, throw, jump, etc). Such classification does not fully eliminate the aforementioned problems of action complexity and ambiguity. Figure 7 shows cases of different atomic action datasets where a single action class contains fundamentally different actions.

HAA150 eliminates any further ambiguity by fine-graining the atomic action (action verb) into target-specific atomic action, e.g., HAA150 fine-grains an English verb "Play (Musical Instrument)" to "Play Cello" or "Play Trumpet", "Throw (an Object)" to "Hammer Throw" or "Shoot Basketball". Our fine-grained atomic actions do not contain any ambiguity, with only a single type of action included under a single class.

4 Experiments

4.1 Human-Centric

To study the effect of our human-centric HAA150, we compare the performance among three models: one trained with just the RGB frames, the other with human poses estimated by [3], and the third model is trained with both RGB and estimated human poses input. Two baseline architectures are used to validate the benefits of our curated atomic action dataset. They are respectively 2D CNN and 3D CNN. In the 2D-model, we adopt ResNet-101 [14]. For the 3D model, we adopt I3D network [4]. Figure 8 shows the model architecture for 3D-based CNNs.

Implementation To train the 2D model, one frame in each training clip is randomly selected. To train the 3D model, we randomly select 32 consecutive frames from a given training clip. In both cases, the training frames are resized with the smallest dimension of 256 pixels using bilinear interpolation with aspect ratio preserved. Pixel values are re-scaled to $[-1, 1]$. During training, we randomly select 224×224 image crops, and select the center 224×224 image crop from the video during testing. The networks are trained with Softmax loss over 150 classes. During testing, all the input frames are used, and the outputs are averaged to get the prediction.

Table 6. Classification results on HAA150

	Top-1	Top-3	# of parameters
ResNet-101 (RGB)	13.32 %	27.55 %	42M
I3D (RGB)	26.64 %	47.45 %	12M
I3D (Human Pose)	41.97 %	62.77 %	12M
I3D (RGB + Human Pose)	**49.82 %**	**72.26** %	12M

Results There are several noteworthy observations. The 3D-based model outperforms the 2D-based models, suggesting that temporal information is indeed a crucial factor for recognizing actions, since multiple actions may share similar poses. The model trained with estimated human pose greatly outperforms the model trained with RGB input only. Note that HAA150 contains scenes from diverse background and environment, which forbids the model from trivially detecting the action by recognizing background scene. This experiment thus validates the benefits of human-centric action.

4.2 Comparison

To show the generalization performance of the models trained with HAA150, we evaluate with a subset of the AVA dataset. We begin by selecting a subset of the

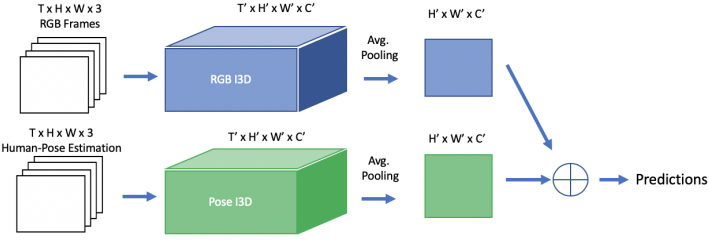

Fig. 8. Model architecture for I3D (RGB + Human Pose). Only a single branch is used respectively for I3D (RGB) and I3D (Human Pose).

Table 7. Results on miniAVA and miniAVA.

	miniAVA	**miniHAA**
Play Instrument	36.8	32.4
Eat	45.9	86.8
Run	86.4	94.6
mAP	56.4	**71.2**

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 based on the fact that these actions are typically mutually exclusive, 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 HAA150 under categories "Run", "Play Instruments", and "Eat", which contains 30 actions to produce **miniHAA**. We train a model with miniHAA and compare the performance with a model trained from scratch with miniAVA and report the mean Average Precision (mAP) over the three classes. Table 7 tabulates the results.

Results Compared to the model trained on miniAVA, the model trained on miniHAA has improved the mAP for 14.8%p. The result seems even more surprising considering the fact that the size of miniHAA ($\sim$ 700) is much smaller than the training set in miniAVA ($\sim$ 7000). The suggests that a class-balanced dataset of human actions in HAA150 helps the model to learn the features in a more efficient manner even HAA150 is much smaller, which is readily extensible to cover more classes with fewer videos required to achieve better performance.

5 Conclusion

This paper introduces HAA150, a new human action dataset with fine-grained atomic action labels and human-centric clip annotations, where the videos are carefully selected (i.e., our collection is "cultivated" rather than "in the wild") such that the relevant human poses are apparent and detectable. With carefully curated action videos, HAA150 does not suffer from temporal noises due to its high sampling frequency on dense action labeling. With a small number of clips per class, HAA150 is highly scalable to include more action classes which is our future work. We have demonstrated the efficacy of HAA150 where action recognition can be greatly benefited from our clean, human-centric and atomic data. Comparison with existing atomic datasets shows that miniHAA performs better than miniAVA on selected action categories, a pleasant surprise which may pose an important next question to answer: does a smaller but clean, human-

centric and atomic action dataset with diverse classes work better than large-scale dataset collected in the wild?

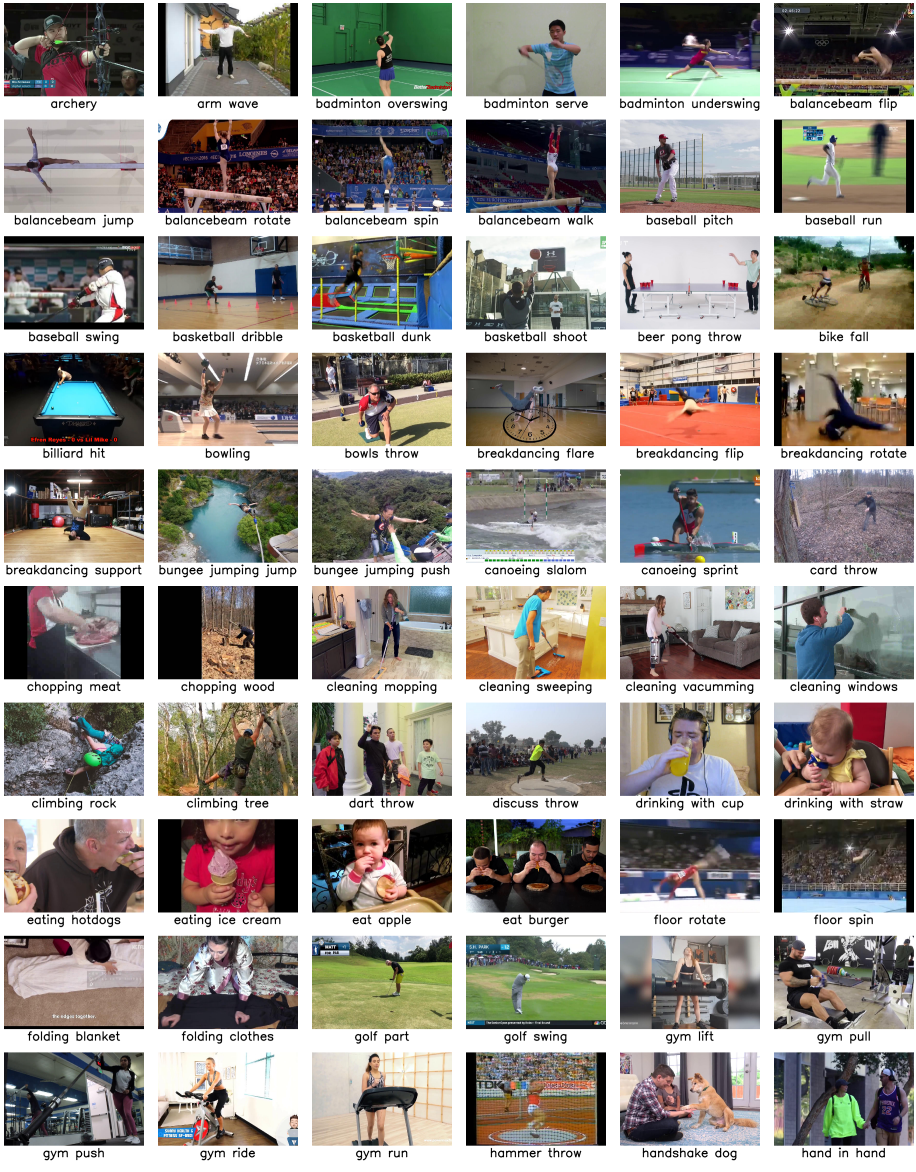

Fig. 9. Sample classes in HAA150

References

1. Baker, R.G.: Midwest and its children: The psychological ecology of an American town. American Journal of Sociology **62**(2) (1956)
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: 2005 IEEE International Conference on Computer Vision (ICCV 2005) (2005)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7291–7299 (2017)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017) (2017)
5. Chaudhuri, S., Roth, J., Ellis, D.P.W., Gallagher, A.C., Kaver, L., Marvin, R., Pantofaru, C., Reale, N., Reid, L.G., Wilson, K.W., Xi, Z.: Ava-speech: A densely labeled dataset of speech activity in movies. In: INTERSPEECH (2018)
6. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: Mars: Motion-augmented rgb stream for action recognition. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019) (2019)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
8. Diba, A., Fayyaz, M., Sharma, V., Arzani, M.M., Yousefzadeh, R., Gall, J., Gool, L.V.: Spatio-temporal channel correlation networks for action classification. In: 2018 European Conference on Computer Vision (ECCV 2018) (2018)
9. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015) (2015)
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. CoRR **abs/1812.03982** (2018), http://arxiv.org/abs/1812.03982
11. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: A better baseline for AVA. CoRR **abs/1807.10066** (2018), http://arxiv.org/abs/1807.10066
12. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The"something something" video database for learning and evaluating visual common sense. In: ICCV. vol. 1, p. 3 (2017)
13. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: AVA: A video dataset of spatio-temporally localized atomic visual actions. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018) (2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016) (2016)
15. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: 2013 IEEE International Conference on Computer Vision (ICCV 2013) (2013)
16. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. In: 2010 International Conference on Machine Learning (ICML 2010) (2010)

17. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J.: STM: spatiotemporal and motion encoding for action recognition. CoRR **abs/1908.02486** (2019), http://arxiv.org/abs/1908.02486
18. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014) (2014)
19. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. CoRR **abs/1705.06950** (2017), http://arxiv.org/abs/1705.06950
20. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: 2005 IEEE International Conference on Computer Vision (ICCV 2005) (2005)
21. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T.A., Serre, T.: HMDB: A large video database for human motion recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV 2011) (2011)
22. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008) (2008)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
24. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009) (2009)
25. Mettes, P., van Gemert, J.C., Snoek, C.G.M.: Spot on: Action localization from pointly-supervised proposals. In: 2016 European Conference on Computer Vision (ECCV 2016) (2016)
26. Monfort, M., Zhou, B., Bargal, S.A., Andonian, A., Yan, T., Ramakrishnan, K., Brown, L.M., Fan, Q., Gutfreund, D., Vondrick, C., Oliva, A.: Moments in time dataset: one million videos for event understanding. CoRR **abs/1801.03150** (2018), http://arxiv.org/abs/1801.03150
27. Niebles, J.C., Wang, H., Li, F.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**(3) (2008)
28. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014) (2014)
29. Pan, B., Sun, J., Lin, W., Wang, L., Lin, W.: Cross-stream selective networks for action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (CVPR 2017) (2019)
30. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: IEEE International Conference on Computer Vision (ICCV 2017) (2017)
31. Rodriguez, M.D., Ahmed, J., Shah, M.: 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)

32. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., Pantofaru, C.: Ava-activespeaker: An audio-visual dataset for active speaker detection. arXiv:1901.01342 (2019)
33. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., Pantofaru, C.: Ava-activespeaker: An audio-visual dataset for active speaker detection (2019)
34. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: 2004 IEEE International Conference on Pattern Recognition (CVPR 2004) (2004)
35. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision (2016)
36. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: 2014 Neural Information Processing Systems (NIPS 2014) (2014), http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos
37. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR **abs/1212.0402** (2012), http://arxiv.org/abs/1212.0402
38. Stroud, J.C., Ross, D.A., Sun, C., Deng, J., Sukthankar, R.: D3D: distilled 3d networks for video action recognition. CoRR **abs/1812.08249** (2018), http://arxiv.org/abs/1812.08249
39. Sun, L., Jia, K., Yeung, D., Shi, B.E.: Human action recognition using factorized spatio-temporal convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV 2015) (2015)
40. Taylor, G.W., Fergus, R., LeCun, Y., Bregler, C.: Convolutional learning of spatio-temporal features. In: 2010 European Conference on Computer Vision (ECCV 2010) (2010)
41. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV 2015) (2015)
42. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018) (2018)
43. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1510–1517 (2018)
44. Weinzaepfel, P., Martin, X., Schmid, C.: Towards weakly-supervised action localization. CoRR **abs/1605.05197** (2016)
45. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: 2018 European Conference on Computer Vision (ECCV 2018) (2018)
46. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009) (2009)
47. Zhao, H., Yan, Z., Torresani, L., Torralba, A.: Hacs: Human action clips and segments dataset for recognition and temporal localization. arXiv preprint arXiv:1712.09374 (2019)
48. Zhou, Y., Sun, X., Zha, Z., Zeng, W.: Mict: Mixed 3d/2d convolutional tube for human action recognition. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018) (2018)

49. Zolfaghari, M., Singh, K., Brox, T.: ECO: efficient convolutional network for online video understanding. In: 2018 European Conference on Computer Vision (ECCV 2018) (2018)