HAA500: Human-Centric Atomic Action Dataset with Curated Videos

Jihoon Chung[1,2] Cheng-hsin Wuu[1,3] Hsuan-ru Yang[1] Yu-Wing Tai[1,4] Chi-Keung Tang[1]
[1]HKUST [2]Princeton University [3]Carnegie Mellon University [4]Kuaishou Technology
jc5933@princeton.edu cwuu@andrew.cmu.edu hyangap@ust.hk yuwing@gmail.com cktang@cs.ust.hk

Abstract

We contribute HAA500[1], a manually annotated human-centric atomic action dataset for action recognition on 500 classes with over 591K labeled frames. To minimize ambiguities in action classification, HAA500 consists of highly diversified classes of fine-grained atomic actions, where only consistent actions fall under the same label, e.g., "Baseball Pitching" vs "Free Throw in Basketball". Thus HAA500 is different from existing atomic action datasets, where coarse-grained atomic actions were labeled with coarse action-verbs such as "Throw". HAA500 has been carefully curated to capture the precise movement of human figures with little class-irrelevant motions or spatio-temporal label noises.

The advantages of HAA500 are fourfold: 1) human-centric actions with a high average of 69.7% detectable joints for the relevant human poses; 2) high scalability since adding a new class can be done under 20–60 minutes; 3) curated videos capturing essential elements of an atomic action without irrelevant frames; 4) fine-grained atomic action classes. Our extensive experiments including cross-data validation using datasets collected in the wild demonstrate the clear benefits of human-centric and atomic characteristics of HAA500, which enable training even a baseline deep learning model to improve prediction by attending to atomic human poses. We detail the HAA500 dataset statistics and collection methodology and compare quantitatively with existing action recognition datasets.

1. Introduction

Observe the *coarse* annotation provided by commonly-used action recognition datasets such as [21, 25, 42], where the same action label was assigned to a given complex video action sequence (*e.g.*, *Play Soccer*, *Play Baseball*) typically lasting 10 seconds or 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 Baseball*).

Recently, atomic action datasets [5, 16, 17, 36, 39] have been introduced in an attempt to resolve the aforementioned issue. Google's AVA actions dataset [17] provides dense annotations of 80 atomic visual actions in 430 fifteen-minute video clips where actions are localized in space and time. AVA spoken activity dataset [36] contains temporally labeled face tracks in videos, where each face instance is labeled as speaking or not, and whether the speech is audible. Something-Something dataset [16] contains clips of humans performing pre-defined basic actions with daily objects.

However, some of their actions are still coarse which can be further split into atomic classes with significantly different motion gestures. *E.g.*, AVA [17] and Something-Something [16] 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 Cello*, and the latter into *Soccer Throw In* and *Pitch Baseball*, *etc*., because each of these atomic actions has significantly different gestures. Encompassing different visual postures into a single 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, ACAR-Net (mAP: 38.30%) [33], in AVA [17], despite only having 80 classes.

The other problem with existing action recognition video datasets is that their training examples contain actions irrelevant to the target action. Video datasets typically have fixed clip lengths, allowing unrelated video frames to be easily included during the data collection stage. Kinetics 400 dataset [21], with a fixed 10-second clip length, contains a lot of irrelevant actions, *e.g.*, showing the audience before the main *violin playing*, or a person takes a long run before *kicking* the ball. Another problem is having too limited or too broad field-of-view, where a video only exhibits a part of a human interacting with an object [16], or a single video contains multiple human figures with different actions present [17, 21, 48].

Recently, FineGym [39] has been introduced to solve the aforementioned limitations by proposing fine-grained action annotations, *e.g.*, *Balance Beam-Dismount-Salto Forward Tucked*. But due to the expensive data collection pro-

[1]HAA500 project page: `https://www.cse.ust.hk/haa`.

This work was supported by Kuaishou Technology and the Research Grant Council of the Hong Kong SAR under grant no. 16201818.

Figure 1. HAA500 is a fine-grained atomic action dataset, with fine-level action annotations (*e.g.*, *Soccer-Dribble*, *Soccer-Throw In*) compared to the traditional composite action annotations (*e.g.*, *Soccer*, *Baseball*). HAA500 is comparable to existing coarse-grained atomic action datasets, where we have distinctions (*e.g.*, *Soccer-Throw In*, *Baseball-Pitch*) within an atomic action (*e.g.*, *Throw Something*) when the action difference is visible. The figure above displays sample videos from three different areas of HAA500. Observe that each video contains one or a few dominant human figures performing the pertinent action.

cess, they only contain 4 events with atomic action annotations (*Balance Beam*, *Floor Exercise*, *Uneven Bars*, and *Vault-Women*), and their clips were extracted from professional gymnasium videos in athletic or competitive events.

In this paper, we contribute Human-centric Atomic Action dataset (**HAA500**) which has been constructed with carefully curated videos with a high average of 69.7% detectable joints, where a dominant human figure is present to perform the labeled action. The curated videos have been annotated with fine-grained labels to avoid ambiguity, and with dense per-frame action labeling and no unrelated frames being included in the collection as well as annotation. HAA500 contains a wide variety of atomic actions, ranging from athletic atomic action (*Figure Skating - Ina Bauer*) to daily atomic action (*Eating a Burger*). HAA500 is also highly scalable, where adding a class takes only 20–60 minutes. 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. Our extensive cross-data experiments validate that precise annotation of fine-grained classes leads to preferable properties against datasets with orders of magnitude larger in size.

Figure 1 shows example atomic actions collected.

2. Related Works

Table 1 summarizes representative action recognition datasets.

2.1. Action Recognition Dataset

Composite Action Dataset Representative action recognition datasets, such as HMDB51 [25], UCF101 [42], Hollywood-2 [29], ActivityNet [9], and Kinetics [3, 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 video classifiers. A few datasets used in action recognition research, such as MSR Actions [47], UCF Sports [34], and JHMDB [19], 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 [42], DALY [44], and Hollywood2Tubes [30] evaluate spatio-temporal localization in untrimmed videos, resulting in a performance drop due to the more difficult nature of the task.

Dataset	Videos	Actions	Atomic
KTH [37]	600	6	✓
Weizmann [2]	90	10	✓
UCF Sports [34]	150	10	
Hollywood-2 [29]	1,707	12	
HMDB51 [25]	7,000	51	
UCF101 [42]	13,320	101	
DALY [44]	510	10	
AVA [17]	387,000	80	✓
Kinetics 700 [3]	650,317	700	
HACS [48]	1,550,000	200	✓
Moments in Time [32]	1,000,000	339	✓
FineGym [39]	32,687	530	✓
HAA500	**10,000**	**500**	✓

Table 1. Summary of representative action recognition datasets.

One common issue on these aforementioned datasets is that they are annotated with composite action classes (*e.g.*, *Playing Tennis*), thus different human action gestures (*e.g.*, *Backhand Swing*, *Forehand Swing*) are annotated under a single class. Another issue is that they tend to capture in wide field-of-view and thus include multiple human figures (*e.g.*, tennis player, referee, audience) with different actions in a single frame, which inevitably introduce confusion to action analysis and recognition.

Atomic Action Dataset To model finer-level events, the AVA dataset [17] was introduced to provide person-centric spatio-temporal annotations on atomic actions similar to some of the earlier works [2, 13, 37]. Other special-

	Kinetics 400 [21]		Something V1 [16]	
Models	Top-1	Top-5	Top-1	Top-5
TSN (R-50) [43]	70.6%	89.2%	20.5%	47.5%
2-Stream I3D [4]	71.6%	90.0%	41.6%	72.2%
TSM (R-50) [27]	74.1%	91.2%	47.3%	76.2%
TPN (TSM) [46]	78.9%	93.9%	50.2%	75.8%
Skeleton-based	Kinetics 400 [21]		NTU-RGB+D [38]	
Models	Top-1	Top-5	X-Sub	X-View
Deep LSTM [38]	16.4%	35.3%	62.9%	70.3%
ST-GCN [45]	30.7%	52.8%	81.5%	88.3%

Table 2. Performance of previous works on Kinetics 400 [21], Something-Something [16], and NTU-RGB+D [38] dataset. We evaluate on both cross-subject (X-Sub) and cross-view (X-View) benchmarks for NTU-RGB+D. For a fair comparison, in this paper we use [21] rather than [3] as representative action recognition model still use [21] for pre-training or benchmarking at the time of writing.

ized datasets such as Moments in Time [32], HACS [48], Something-Something [16], and Charades-Ego [40] provide classes for atomic actions but none of them is a human-centric atomic action, where some of the video frames are ego-centric which only show part of a human body (*e.g.*, hand), or no human action at all. Existing atomic action datasets [17, 32] tend to have atomicity under English linguistics, *e.g.*, in Moments in Time [32] *Open* is annotated on video clips with a tulip opening, an eye opening, a person opening a door, or a person opening a package, which is fundamentally different actions only sharing the verb *open*, which gives the possibility of finer division.

Fine-Grained Action Dataset Fine-grained action datasets try to solve ambiguous temporal annotation problems that were discussed in [1, 31]. These datasets (*e.g.*, [6, 14, 24, 26, 35, 39]) use systematic action labeling to annotate fine-grained labels on a small domain of actions. Breakfast [24], MPII Cooking 2 [35], and EPIC-KITCHENS [6] offer fine-grained actions for cooking and preparing dishes, *e.g.*, *Twist Milk Bottle Cap* [24]. JIGSAWS [14], Diving48 [26], and FineGym [39] offer fine-grained action datasets respectively for surgery, diving, and gymnastics. While existing fine-grained action datasets are well suited for benchmarks, due to their low variety and the narrow domain of the classes, they cannot be extended easily in general-purpose action recognition.

Our HAA500 dataset differs from all of the aforementioned datasets as we provide a wide variety of 500 fine-grained atomic human action classes in various domains, where videos in each class only exhibit the relevant human atomic actions.

2.2. Action Recognition Architectures

Current action recognition architectures can be categorized into two major approaches: 2D-CNN and 3D-CNN. 2D-CNN [8, 12, 27, 41, 43, 49] based models utilize image-based 2D-CNN models on a single frame where features are aggregated to predict the action. While some methods (*e.g.*, [8]) use RNN modules for temporal aggregation over visual features, TSN [43] shows that simple average pooling can be an effective method to cope with temporal aggregation. To incorporate temporal information to 2D-CNN, a two-stream structure [12, 41] has been proposed to use RGB-frames and optical flow as separate inputs to convolutional networks. 3D-CNN [4, 11, 20] takes a more natural approach by incorporating spatio-temporal filters into the image frames. Inspired from [41], two-streamed inflated 3D-CNN (I3D) [4] incorporates two-stream structure on 3D-CNN. SlowFast [11] improves from I3D by showing that the accuracy increases when the 3D kernels are used only in the later layers of the model. A different approach is adopted in TPN [46] where a high-level structure is designed to adopt a temporal pyramid network which can use either 2D-CNN or 3D-CNN as a backbone. Some models [22, 23, 45] use alternative information to predict video action. Specifically, ST-GCN [45] uses a graph convolutional network to predict video action from pose estimation. However, their pose-based models cannot demonstrate better performance than RGB-frame-based models.

Table 2 tabulates the performance of representative action recognition models on video action datasets, where 2D-skeleton based models [38, 45] show considerably low accuracy in Kinetics 400 [21].

3. HAA500

3.1. Data Collection

The annotation of HAA500 consists of two stages: vocabulary collection and video clip selection. While the bottom-up approach which annotates action labels on selected long videos was often used in atomic/fine-grained action datasets [17, 39], we aim to build a clean and fine-grained dataset for atomic action recognition, thus the video clips are collected based on pre-defined atomic actions following a top-down approach.

3.1.1 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 4 areas; *Sport/Athletics*, *Playing Musical Instruments*, *Games and Hobbies*, and *Daily Actions*, where future extension beyond the existing classes is feasible. We select action labels where the variations within a class are typically indistinguishable. For example, instead of *Hand Whistling*, we have *Whistling with One Hand* and *Whistling with Two Hands*, as the variation is large and distinguishable. Our vocabulary collection methodology makes the dataset hierarchical where atomic actions may be combined to form a composite action, *e.g.*, *Whistling* or *Playing Soccer*. Consequently, HAA500 contains 500 atomic action classes, where 212 are *Sport/Athletics*, 51 are *Playing Musical Instruments*, 82 are *Games and Hobbies* and 155 are *Daily Actions*.

(a) Kinetics 400 - Shooting Basketball
(b) Kinetics 400 - Singing
(c) HACS - Long Jump
(d) HAA500 - Uneven Bars: Land

Figure 2. Different types of label noise in action recognition datasets. **(a)**: Kinetics 400 has a 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)**: A camera cut can be seen, showing unrelated frames (audience) after the main action. **(c)**: By not having a frame-accurate clipping, the clip starts with a person-of-interest in the midair, and quickly disappears after few frames, causing the rest of the video clip not to have any person in action. **(d)**: Our HAA500 accurately annotates the full motion of *Uneven Bars - Land* without any irrelevant frames. All the videos in the class start with the exact frame an athlete puts the hand off the bar, to the exact frame when he/she finishes the landing pose.

action	clips	mean length	duration	frames
500	10,000	2.12s	21,207s	591K

no. of people	1	2	>2
	8,309	859	832

moving camera	O	X
	2,373	7,627

Table 3. Summary of HAA500.

3.1.2 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 and each class of atomic action containing 16 training clips. We manually select the clips with apparent human-centric actions where the person-of-interest is the only dominant person in the frame at the center with their body clearly visible. To increase diversity among the video clips and avoid unwanted bias, all the clips were collected from different YouTube videos, with different environment settings so that the action recognition task cannot be trivially reduced to identifying the corresponding backgrounds. Clips are properly trimmed in a frame-accurate manner to cover the desired actions while assuring every video clip to have compatible actions within each class (*e.g.*, every video in the class *Salute* starts on the exact frame where the person is standing still before moving the arm, and the video ends when the hand goes next to the eyebrow). Refer to Figure 1 again for examples of the selected videos.

3.1.3 Statistics

Table 3 summarizes the HAA500 statistics. HAA500 includes 500 atomic action classes where each class contains 20 clips, with an average length of 2.12 seconds. Each clip was annotated with meta-information which contains the following two fields: the number of dominant people in the video and the camera movement.

Dataset	Clip Length	Irr. Actions	Camera Cuts
UCF101 [42]	Varies		
HMDB51 [25]	Varies		✓
AVA [17]	1 second	✓	✓
HACS [48]	2 second	✓	
Kinetics [21]	10 second	✓	✓
M.i.T. [32]	3 second		
HAA500	Just Right		

Table 4. Clip length and irrelevant frames of video action datasets.

3.1.4 Training/Validation/Test Sets

Since the clips in different classes are mutually exclusive, all clips appear only in one split. The 10,000 clips are split as 16:1:3, resulting in segments of 8,000 training, 500 validation, and 1,500 test clips.

3.2. Properties and Comparison

3.2.1 Clean Labels for Every Frame

Most video datasets [17, 21, 42] show strong label noises, due to the difficulties of collecting clean video action datasets. Some [21, 25, 42] often focus on the “scene” of the video clip, neglecting the human “action” thus including irrelevant actions or frames with visible camera cuts in the clip. Also, video action datasets [17, 21, 32, 48] have fixed-length video clips, so irrelevant frames are inevitable for shorter actions. Our properly trimmed video collection guarantees a clean label for every frame.

Table 4 tabulates clip lengths and label noises of video action datasets. Figure 2 shows examples of label noises. As HAA500 is constructed with accurate temporal annotation in mind, we are almost free from any adverse effects due to these noises.

3.2.2 Human-Centric

One potential problem in action recognition is that the neural network may predict by trivially comparing the background scene in the video, or detecting key elements in a

Figure 3. The video clips in AVA, HACS, and Kinetics 400 contain multiple human figures 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 HAA500 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.

Dataset	Detectable Joints
Kinetics 400 [21]	41.0%
UCF101 [42]	37.8%
HMDB51 [25]	41.8%
FineGym [39]	44.7%
HAA500	**69.7%**

Table 5. Detectable joints of video action datasets. We use AlphaPose [10] to detect the largest person in the frame, and count the number of joints with a score higher than 0.5.

frame (*e.g.*, a basketball to detect *Playing Basketball*) rather than recognizing the pertinent human gesture, thus causing the action recognition to have no better performance improvements over scene/object recognition. The other problem stems from the video action datasets where videos captured in wide field-of-view contain multiple people in a single frame [17, 21, 48], while videos captured using narrow field-of-view only exhibit very little body part in interaction with the pertinent object [16, 32].

In [17] 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 3 illustrates example frames of different video action datasets.

HAA500 contributes a curated dataset where human joints can be clearly detected over any given frame, thus allowing the model to benefit from learning human movements than just performing scene recognition. As tabulated in Table 5, HAA500 has high detectable joints [10] of 69.7%, well above other representative action datasets.

3.2.3 Atomic

Existing atomic action datasets such as [5, 17, 32] are limited by English linguistics, where action verbs (*e.g.*, walk, throw, pull, *etc.*) are decomposed. Such classification does not fully eliminate the aforementioned problems of composite action datasets. Figure 4 shows cases of different atomic action datasets where a single action class contains fundamentally different actions.

On the other hand, our fine-grained atomic actions contain only a single type of action under each class, *e.g.*, *Baseball - Pitch*, *Yoga - Tree*, *Hopscotch - Spin*, *etc.*

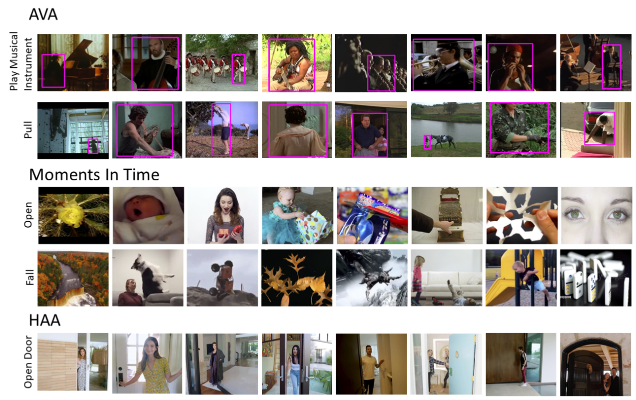

Figure 4. Coarse-grained atomic action datasets label different actions under a single English action verb. HAA500 (Bottom) has fine-grained classes where the action ambiguities are eliminated as much as possible.

3.2.4 Scalability

Requiring only 20 video annotations per class, or around 600 frames to characterize a human-centric atomic action curated as described above, our class-balanced dataset is highly scalable compared to other representative datasets requiring annotation of hundreds or even thousands of videos. In practice, our annotation per class takes around 20–60 minutes including searching the Internet for videos with expected quality. The detailed annotation procedure is available in the supplementary material.

4. Empirical Studies

We study HAA500 over multiple aspects using widely used action recognition models. Left of Table 6 shows the performance of the respective models when they are trained with HAA500. For a fair comparison between different models and training datasets, all the experiments have been performed using hyper parameters given by the original authors without ImageNet [7] pre-training.

For Pose models except for ST-GCN [45], we use three-channel pose joint heatmaps [10] to train pose models. RGB, Flow [18] and Pose [10] all show relatively similar performance in HAA500, where none of them shows superior performance than the others. Given that pose heatmap has far less information than given from RGB frames or optical flow frames, we expect that easily detectable joints of HAA500 benefit the pose-based model performance.

Model		500 Atomic		Inst.	Inst. with Atomic	Sport	Sport with Atomic
		Top-1	Top-3	Top-1	Top-1	Top-1	Top-1
I3D [4]	RGB	33.53%	53.00%	70.59%	**71.90%**	47.48%	**53.93%**
	Flow	34.73%	52.40%	73.20%	**77.79%**	51.42%	**54.40%**
	Pose	35.73%	54.07%	69.28%	**71.90%**	54.87%	55.03 %
	Three-Stream	49.87%	66.60%	81.70%	82.35%	68.55%	**69.81%**
SlowFast [11]	RGB	25.07%	44.07%	40.52%	**50.98%**	42.92%	**44.18%**
	Flow	22.87%	36.93%	71.90%	71.90 %	44.81%	**45.91%**
	Pose	28.33%	45.20%	64.71%	**66.01%**	42.45%	**50.00%**
	Three-Stream	39.93%	56.00%	67.97%	**73.86%**	59.91%	**62.89%**
TSN [43]	RGB	55.33%	75.00%	**86.93%**	84.31%	72.64%	72.48%
	Flow	49.13%	66.60%	79.08%	**86.27%**	**69.97%**	68.24%
	Two-Stream	64.40%	80.13%	89.54%	90.20%	**81.13%**	78.93%
TPN [46]	RGB	50.53%	68.13%	73.20%	**75.82%**	61.64%	**64.15%**
ST-GCN [45]	Pose	29.67%	47.13%	67.32%	67.97%	40.25%	**43.87%**

Table 6. **Left**: HAA500 trained over different models. **Right**: Composite action classification accuracy of different models when they are trained with/without atomic action classification. Numbers are bolded when the difference is larger than 1%.

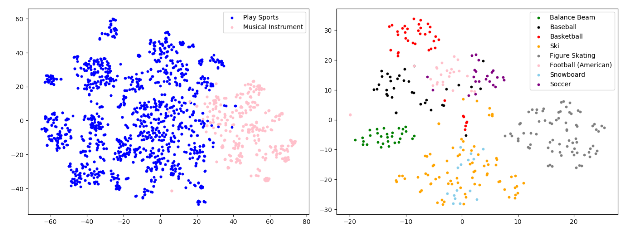

Figure 5. Visualization of HAA500. We extract 1024-vectors from the second last layer of RGB-I3D and plot them using t-SNE.

Furthermore, we study the benefits of atomic action annotation on video recognition, as well as the importance of human-centric characteristics of HAA500. In this paper, we use I3D-RGB [4] with 32 frames for all of our experiments unless otherwise specified. We use AlphaPose [10] for the models that require human pose estimation.

4.1. Visualization

To study the atomic action recognition, we train RGB-I3D model on HAA500 and extract embedding vectors from the second last layer and plot them using truncated SVD and t-SNE. From Figure 5, the embedding vectors show clear similarities to the natural hierarchy of human action. On the left of the figure, we see a clear distinction between classes in *Playing Sports* and classes in *Playing Musical Instruments*. Specifically, in sports, we see similar super-classes, *Snowboarding* and *Skiing*, under close embedding space, while *Basketball*, *Balance Beam (Gymnastics)*, and *Figure Skating* are in their distinctive independent spaces. We observe super-class clustering of composite actions when only the atomic action labeling has been used to train the model. This visualization hints the benefit of fine-grained atomic action labeling for composite action classification tasks.

4.2. Atomic Action

We have previously discussed that modern action recognition datasets introduce ambiguities where two or more composite actions sharing the same atomic actions, while a single composite action class may contain multiple distinguishable actions (*e.g.*, a composite action *Playing Soccer* has *Soccer-Dribble*, *Soccer-Throw*, *etc.*). HAA500 addresses this issue by providing fine-grained atomic action labels that distinguish similar atomic action in different composite actions.

To study the benefits of atomic action labels, specifically, how it helps composite action classification for ambiguous classes, we selected two areas from HAA500, *Sports/Athletics* and *Playing Musical Instruments*, in which composite actions contain strong ambiguities with other actions in the area. We compare models trained with two different types of labels: 1) only composite labels and 2) atomic + composite labels, then we evaluate the performance on composite action classification. Results are tabulated on the right of Table 6. Accuracy of the models trained with only composite labels are under *Inst.* and *Sport* column, and the accuracy of composite action classification trained with atomic action classification is listed on the other columns.

We can observe improvements in composite action classification when atomic action classification is incorporated. The fine-grained action decomposition in HAA500 enables the models to resolve ambiguities of similar atomic actions and helps the model to learn the subtle differences in the atomic actions across different composite actions. This demonstrates the importance of proper labeling of fine-grained atomic action which can increase the performance for composite action classification without changing the model architecture or the training set.

4.3. Human-Centric

HAA500 is designed to contain action clips with a high percentage of detectable human figures. To study the importance of human-pose in fine-grained atomic action recognition, we compare the performance of HAA500 and FineGym when both RGB and pose estimation are given as in-

	RGB	Pose	RGB + Pose
HAA500	33.53%	35.73%	42.80%
Sport	38.52%	47.33%	50.94%
Instrument	30.72%	24.18%	32.03%
Hobbies	31.30%	26.42%	35.37%
Daily	28.82%	28.60%	39.14%
Gym288 [39]	76.11%	65.16%	77.31%

Table 7. Atomic action classification accuracy when both RGB image and pose estimation are given as an input. We also show performance when they are trained separately for comparison.

	UCF101 [42]	ActNet 100 [9]	HMDB51 [25]
Pre-trained	Top-1	Top-1	Top-1
None	58.87%	43.54%	28.56%
AVA [17]	48.54%	30.51%	25.28%
Gym288 [39]	**69.94%**	43.79%	36.24%
UCF101 [42]	-	42.94%	32.37%
ActNet 100 [9]	57.52%	-	28.63%
HMDB51 [25]	53.36%	39.33%	-
HAA500	68.70%	**47.75%**	**40.45%**
Relaxed	62.24%	38.30%	33.29%

Table 8. Fine-tuning performance on I3D.

put. For pose estimation, we obtain the 17 joint heatmaps from AlphaPose [10] and merge them into 3 channels; head, upper-body, and lower-body.

Table 7 tabulates the results. In three out of four areas of HAA500, I3D-RGB shows better performance than I3D-Pose, due to the vast amount of information given to the model. I3D-Pose shows the highest performance on *Sports/Athletics* with vibrant and distinctive action, while I3D-Pose fails to show comparable performance in *Playing Musical Instrument* area, where predicting the atomic action from only 17 joints is quite challenging. Nonetheless, our experiments show a performance boost when both pose estimation and RGB frame are fed to the atomic action classification model, implicating the importance of human action in HAA500 action classification. For FineGym - Gym288, due to the rapid athletic movements resulting in blurred frames, the human pose is not easily recognizable which accounts for relatively insignificant improvements when pose has been used.

5. Observations

We present notable characteristics observed from HAA500 with our cross-dataset experiments.

Effects of Fine-Tuning over HAA500 Here, we test how to exploit the curated HAA500 dataset to detect action in "in-the-wild" action datasets. We pre-train I3D-RGB [4] using HAA500 or other video action datasets [9, 17, 25, 39, 42], and freeze all the layers except for the last three for feature extraction. We then fine-tune the last three layers with "in-the-wild" composite action datasets [9, 25, 42].

Table 8 tabulates the fine-tuning result. Our dataset is carefully curated to have a high variety of backgrounds and

	Original		Normalized	
	Composite	Both	Composite	Both
I3D-RGB	66.01%	56.86%	**75.82%**	**77.12%**
I3D-Flow	73.20%	**77.78%**	**75.16%**	74.51%
2-Stream	77.78%	80.39%	**83.01%**	80.39%

Table 9. Accuracy improvements on person-of-interest normalization. Numbers are composite action classification accuracy.

people while having consistent actions over each class. Despite being comparably smaller and more "human-centric" than other action recognition datasets, HAA500's cleanness and high variety make it easily transferable to different tasks and datasets.

Effects of Scale Normalization HAA500 has high diversity in human positions across the video collection. Here, we choose an area of HAA500, *Playing Musical Instruments*, to investigate the effect of human-figure normalization on detection accuracy. We have manually annotated the bounding box of the person-of-interest in each frame and cropped them for the model to focus on the human action. In Table 9, we test models that were trained to detect the composite actions or both composite and atomic actions.

While HAA500 is highly human-centric with person-of-interest as the most dominant figure of the frame, action classification on the normalized frames still shows considerable improvement when trained on either atomic action annotations or composite action annotations. This indicates the importance of spatial annotation for action recognition.

Effects of Object Detection In most video action datasets, non-human objects exist as a strong bias to the classes (*e.g.*, basketball in *Playing Basketball*). When highly diverse actions (*e.g.*, *Shooting a Basketball*, *Dribbling a Basketball*, *etc*.) are annotated under a single class, straightforward deep-learning models tend to suffer from the bias and will learn to detect the easiest common factor (basketball) among the video clips, rather than "seeing" the pertinent human action. Poorly designed video action dataset encourages the action classification model to trivially become an object detection model.

In HAA500, every video clip in the same class contains compatible actions, making the common factor to be the "action", while objects are regarded as "ambiguities" that spread among different classes (*e.g.*, basketball exists in both *Shooting a Basketball* and *Dribbling a Basketball*). To test the influence of "object" in HAA500, we design an experiment similar to investigating the effect of human poses, as presented in Table 7, where we use object detection heatmap instead. Here we use Fast RCNN [15] trained with COCO [28] dataset to generate the object heatmap. Among 80 detectable objects in COCO, we select 42 objects in 5 categories (sports equipment, food, animals, cutleries, and vehicles) to draw a 5-channel heatmap. Similar to Table 7, the heatmap channel is appended to the RGB channel as input.

	RGB	+ Object
HAA500	33.53%	33.73%
Sport	38.52%	38.68%
Instrument	30.72%	30.07%
HAA-COCO	34.26%	34.26%
UCF101	57.65%	60.19%

Table 10. Accuracy of I3D when trained with object heatmap. HAA-COCO denotes 147 classes of HAA500 expected to have objects that were detected.

Table 10 tabulates the negligible effect of objects in atomic action classification of HAA500, including the classes that are expected to use the selected objects (HAA-COCO), while UCF101 shows improvements when object heatmap is used as a visual cue. Given the negligible effect of object heatmaps, we believe that fine-grained annotation of actions can effectively eliminate unwanted ambiguities or bias ("objects") while in UCF101 (composite action dataset), "objects" can still affect action prediction.

Effects of Dense Temporal Sampling The top of Table 11 tabulates the performance difference of HAA500 and other datasets over the number of frames used during training and testing. The bottom of Table 11 tabulates the performance with varying strides with a window size of 32 frames, except AVA which we test with 16 frames. Top-1 accuracies on action recognition are shown except AVA which shows mIOU due to its multi-labeled nature of the dataset.

As expected, most datasets show the best performance when 32 frames are fed. AVA shows a drop in performance due to the irrelevant frames (*e.g*., action changes, camera cuts, *etc*.) included in the wider window. While all the datasets show comparable accuracy when the model only uses a single frame (*i.e*., when the problem has been reduced to a "Scene Recognition" problem), both HAA500 and Gym288 show a significant drop compared to their accuracy in 32 frames. While having an identical background contributes to the performance difference for Gym288, from HAA500, we see how temporal action movements are crucial for the detection of atomic actions, and they cannot be trivially detected using a simple scene detecting model.

We also see that the density of the temporal window is another important factor in atomic action classification. We see that both HAA500 and Gym288, which are fine-grained action datasets, show larger performance drops when the frames have been sampled with strides of 2 or more, reflecting the importance of sampling for short temporal action movements in fine-grained action classification.

Quality versus Quantity To study the importance of our precise temporal annotation against the size of a dataset, we modify HAA500 by relaxing the temporal annotation requirement, *i.e*., we take a longer clip than the original annotation. Our relaxed-HAA500 consists of 4400K labeled frames, a significant increase from the original HAA500 with 591K frames. Table 12 tabulates the performance comparison between the original and the relaxed version of HAA500 on the original HAA500 test set. We observe the performance drop in all areas, with a significant drop in *Playing Sports*, where accurate temporal annotation benefits the most. Performance drop in *Playing Musical Instruments* area is less significant, as start/finish of action is vaguely defined in these classes. We also test the finetuning performance of relaxed-HAA500, where the bottommost row of Table 8 tabulates the performance drop when the relaxed-HAA500 is used for pre-training. Both of our experiments show the importance of accurate temporal labeling over the size of a dataset.

# of frames	HAA500	UCF101 [42]	AVA [17]	Gym288 [39]
1	19.93%	45.57%	33.57%	39.77%
2	23.27%	47.26%	39.42%	44.68%
4	24.40%	49.30%	39.48%	51.22%
8	24.07%	49.80%	42.38%	59.64%
16	28.20%	52.31%	43.11%	69.25%
32	33.53%	57.65%	29.88%	76.11%
stride 2	27.47%	57.23%	41.49%	68.68%
stride 4	23.87%	52.29%	40.52%	60.76%
stride 8	18.47%	47.95%	38.45%	39.31%

Table 11. Performance comparison on I3D-RGB over the number of frames and strides, wherein the latter a window size of 32 frames is used except AVA which we test with 16 frames.

	HAA500	Relaxed
Overall	**33.53%**	22.80%
Sport	**38.52%**	25.47%
Instrument	**30.72%**	28.10%
Hobbies	**31.30%**	20.33%
Daily	**28.82%**	18.71%

Table 12. Action classification accuracy of original HAA500 and the relaxed version.

6. Conclusion

This paper introduces HAA500, a new human action dataset with fine-grained atomic action labels and human-centric clip annotations, where the videos are carefully selected such that the relevant human poses are apparent and detectable. With carefully curated action videos, HAA500 does not suffer from irrelevant frames, where videos clips only exhibit the annotated action. With a small number of clips per class, HAA500 is highly scalable to include more action classes. We have demonstrated the efficacy of HAA500 where action recognition can be greatly benefited from our clean, highly diversified, class-balanced fine-grained atomic action dataset which is human-centric with a high percentage of detectable poses. On top of HAA500, we have also empirically investigated several important factors that can affect the performance of action recognition. We hope HAA500 and our findings could facilitate new advances in video action recognition.

References

[1] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *ECCV 2018*.

[2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICCV 2005*.

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR 2017*.

[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.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV 2018*.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*.

[8] 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 *CVPR 2015*.

[9] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR 2015*.

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV 2017*.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR 2019*.

[12] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR 2016*.

[13] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Temporal localization of actions with actoms. *TPAMI 2013*.

[14] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *Miccai workshop: M2cai 2014*.

[15] Ross Girshick. Fast r-cnn. In *CVPR 2015*.

[16] 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 2017*.

[17] 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 *CVPR 2018*.

[18] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR 2017*.

[19] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *ICCV 2013*.

[20] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. STM: Spatiotemporal and motion encoding for action recognition. In *ICCV 2019*.

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

[22] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR 2017*.

[23] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *CVPRW 2017*.

[24] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR 2014*.

[25] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV 2011*.

[26] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV 2018*.

[27] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *ICCV 2019*.

[28] 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 *ECCV 2014*.

[29] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *CVPR 2009*.

[30] Pascal Mettes, Jan C. van Gemert, and Cees G. M. Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV 2016*.

[31] Davide Moltisanti, Michael Wray, Walterio Mayol-Cuevas, and Dima Damen. Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video. In *ICCV 2017*.

[32] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI 2019*.

[33] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *CVPR 2021*.

[34] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR 2008*.

[35] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *IJCV 2016*.

[36] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*.

[37] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *CVPR 2004*.

[38] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *CVPR 2016*.

[39] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR 2020*.

[40] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV 2016*.

[41] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS 2014*.

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

[43] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *TPAMI 2018*.

[44] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Towards weakly-supervised action localization. *arXiv preprint arXiv:1605.05197*, 2, 2016.

[45] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI 2018*.

[46] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR 2020*.

[47] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative subvolume search for efficient action detection. In *CVPR 2009*.

[48] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV 2019*.

[49] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV 2018*.

HAA500: Supplementary Material

Jihoon Chung[1,2] Cheng-hsin Wuu[1,3] Hsuan-ru Yang[1] Yu-Wing Tai[1,4] Chi-Keung Tang[1]
[1]HKUST [2] Princeton University [3]Carnegie Mellon University [4]Kuaishou Technology
jc5933@princeton.edu cwuu@andrew.cmu.edu hyangap@ust.hk yuwing@gmail.com cktang@cs.ust.hk

1. Video Collection Procedure

To guarantee a clean dataset with no label noises, we adopt a strict video collecting methodology for every class. We detail the method below.

1. We assign a single annotator for a single class. This is to assure that the same rule applies to every video in a class.
2. The action class is classified as either continuous action or discrete action. Discrete action is when the action can have a single distinguishable action sequence. (*e.g.*, *Baseball-Swing*, *Yoga-Bridge*, etc.). Continuous action otherwise. (*Running*, *Playing Violin*, etc.)
   (a) If it is discrete, make an internal rule to define the action. (e.g., *Jumping Jack* starts and ends when the person is standing still. The video clip contains only a single jump. *Push-up* starts and ends when the person is at the highest point. It should only have a single push-up). Every video should follow the internal rule so that every action in the class has compatible motion.
   (b) For continuous, we take video clips with appropriate length.
3. Here are rules that the annotator has to follow.
   - 20 videos should be unique to each other with a varied person, varied backgrounds.
   - The person in action should be the dominant person of the frame. If there are people of non-interest, they should not be performing any action.
   - Camera cuts should not exist.
   - Every video should include a large portion of the human body.
   - It is fine to have action variance that doesn't influence the semantics of the action. (*e.g.*, a person can sit or stand in *Whistling with One Hand* as long as the motion of whistling exists.)
   - 20 videos are split into train/val/test set by 16/1/3. The validation set contains the "standard" body action of the class, and 3 videos in the test set should be well diverse.
4. Two or more reviewers that are not the annotator review the video to check for any mistakes.

2. Experiment Detail

In this section, we explain some of the experiment details of our paper.

Variable Length of a Video For model [**?**, **?**, **?**, **?**], we randomly select 32 adjacent frames of a video during training. If the video is shorter than 32 frames, we replicate the last frame to match the size. During testing, we replicate the last frame to match the size to a multiple of 32, where the video is then divided into smaller mini-clips of size 32. The prediction score of each mini-clip is averaged to get the final prediction. In Table 11, where we train with fewer frames, we zero-pad on both ends to size 16. On ST-GCN [**?**] we follow the same procedure of the original paper, where the video is either truncated or replicated to match the length of 300.

Implementation In all of our experiments, we use PyTorch for our deep learning framework. We use the official code of the model when they are available. While we use the same hyperparameters which the authors used for their model, for a fair comparison we do not pre-train the model before training.

3. List of Classes in HAA500

Here, we list classes of HAA500 in each area.

Sports/Athletics

1. Abseiling
2. Archery
3. Backflip
4. Backward Roll

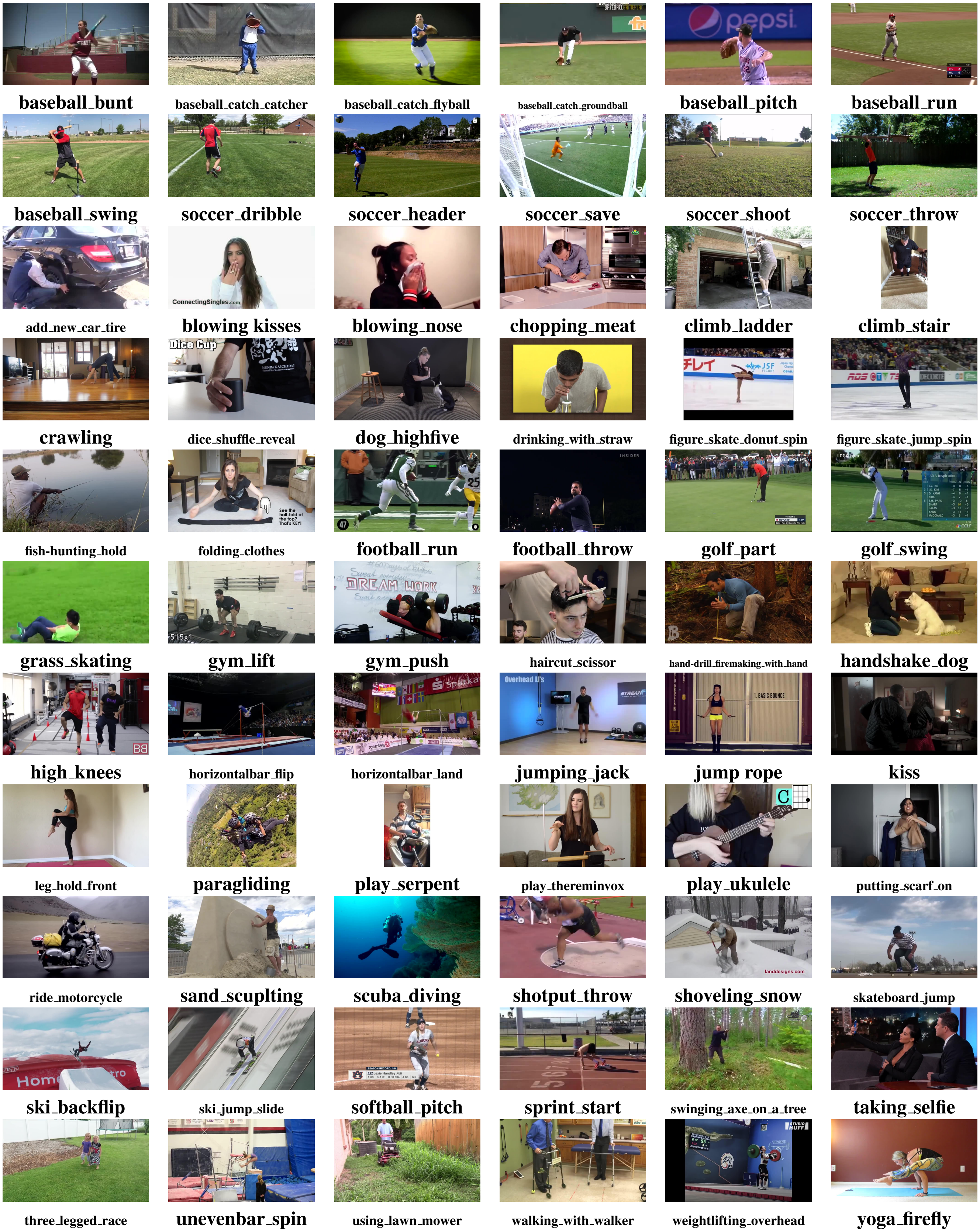

Figure 1. Video samples of different classes.

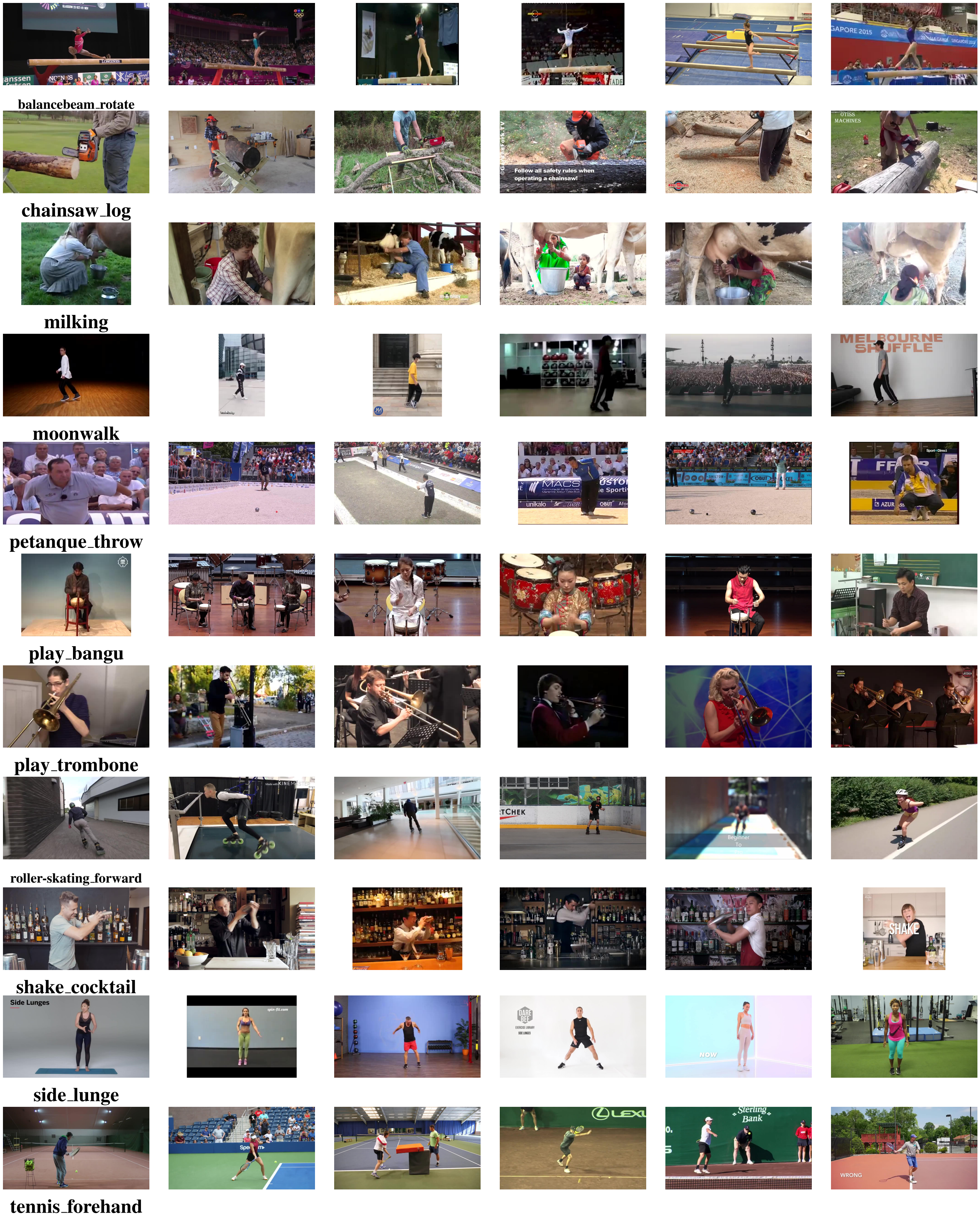

Figure 2. HAA500 contains diverse videos per action class.

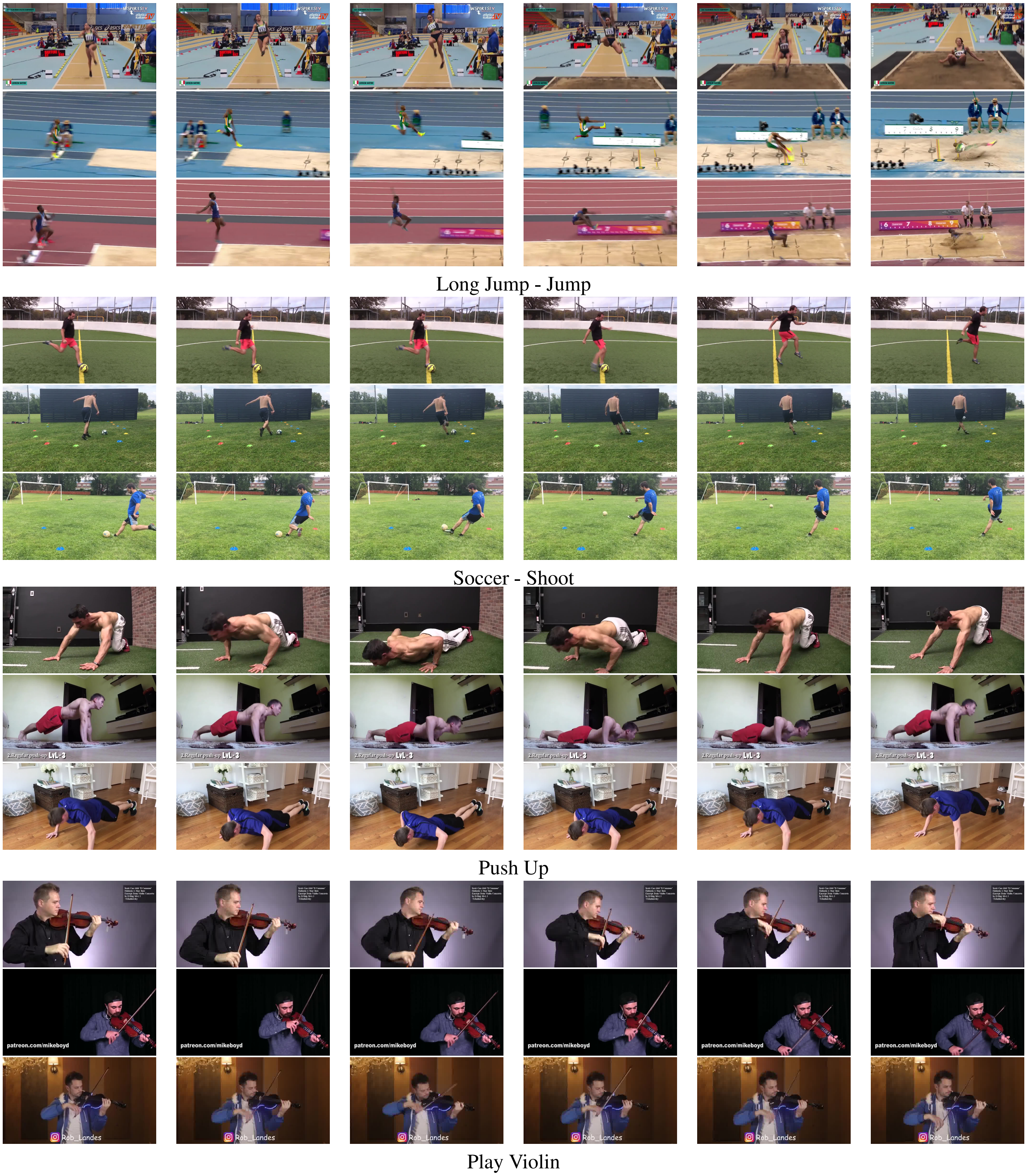

Figure 3. Six sample frames of different videos. Each frame has an equal distance from the other, the first and the last sample frame are the first and the last frame of the video. In discrete action classes, (*Long Jump - Jump*, *Push Up*, *Soccer - Shoot*), every video in the class shows a single motion. For action classes where it is hard to define a single motion (*i.e.*, continuous actions, *e.g.*, *Play Violin*), videos are cut in appropriate length.

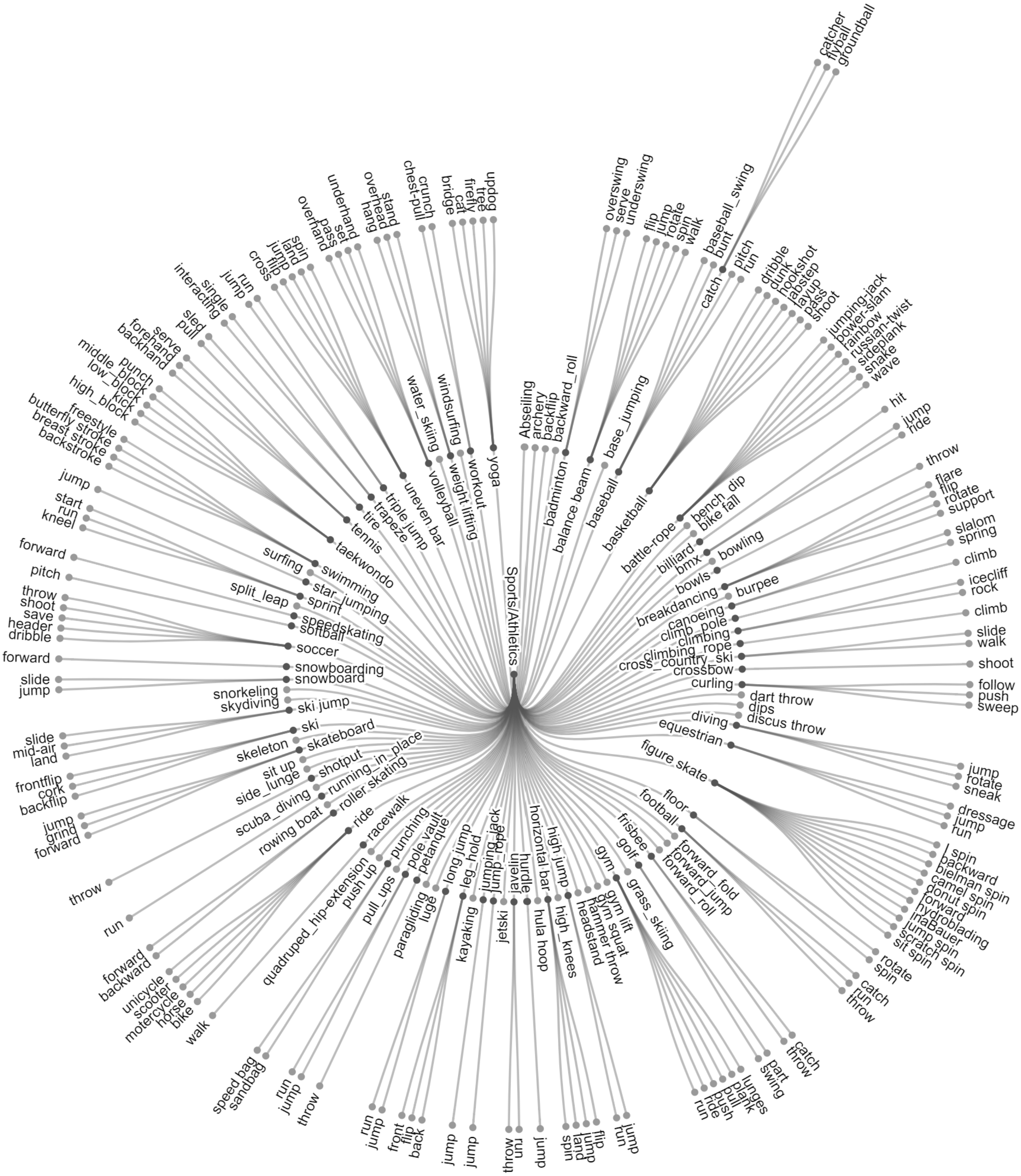

Figure 4. Hierarchy of action classes in Sports/Athletics area.

5. Badminton Overswing
6. Badminton Serve
7. Badminton Underswing
8. Balance Beam Flip
9. Balance Beam Jump
10. Balance Beam Rotate
11. Balance Beam Spin
12. Balance Beam Walk

13. Base Jumping
14. Baseball Baseball Swing
15. Baseball Bunt
16. Baseball Pitch
17. Baseball Run
18. Basketball Dribble
19. Basketball Dunk
20. Basketball Hookshot
21. Basketball Jabstep
22. Basketball Layup
23. Basketball Pass
24. Basketball Shoot
25. Battle-Rope Jumping-Jack
26. Battle-Rope Power-Slam
27. Battle-Rope Rainbow
28. Battle-Rope Russian-Twist
29. Battle-Rope Sideplank
30. Battle-Rope Snake
31. Battle-Rope Wave
32. Bench Dip
33. Bike Fall
34. Billiard Hit
35. Bmx Jump
36. Bmx Ride
37. Bowling
38. Bowls Throw
39. Breakdancing Flare
40. Breakdancing Flip
41. Breakdancing Rotate
42. Breakdancing Support
43. Burpee
44. Canoeing Slalom
45. Canoeing Spring
46. Catch Catcher
47. Catch Flyball
48. Catch Groundball
49. Climb Pole Climb
50. Climbing Icecliff
51. Climbing Rock
52. Climbing Rope Climb
53. Cross Country Ski Slide
54. Cross Country Ski Walk
55. Crossbow Shoot
56. Curling Follow
57. Curling Push
58. Curling Sweep
59. Dart Throw
60. Dips
61. Discus Throw
62. Diving Jump
63. Diving Rotate
64. Diving Sneak
65. Equestrian Dressage
66. Equestrian Jump
67. Equestrian Run
68. Figure Skate I Spin
69. Figure Skate Backward
70. Figure Skate Bielman Spin
71. Figure Skate Camel Spin
72. Figure Skate Donut Spin
73. Figure Skate Forward
74. Figure Skate Hydroblading
75. Figure Skate Inabauer
76. Figure Skate Jump Spin
77. Figure Skate Scratch Spin
78. Figure Skate Sit Spin
79. Floor Rotate
80. Floor Spin
81. Football Catch
82. Football Run
83. Football Throw
84. Forward Fold
85. Forward Jump
86. Forward Roll
87. Frisbee Catch
88. Frisbee Throw
89. Golf Part
90. Golf Swing
91. Grass Skiing
92. Gym Lift
93. Gym Lunges
94. Gym Plank
95. Gym Pull
96. Gym Push
97. Gym Ride
98. Gym Run
99. Gym Squat
100. Hammer Throw
101. Headstand
102. High Jump Jump
103. High Jump Run
104. High Knees
105. Horizontal Bar Flip
106. Horizontal Bar Jump
107. Horizontal Bar Land
108. Horizontal Bar Spin
109. Hula Hoop
110. Hurdle Jump
111. Javelin Run
112. Javelin Throw
113. Jetski
114. Jump Rope Jump
115. Jumping Jack Jump
116. Kayaking
117. Leg Hold Back
118. Leg Hold Flip
119. Leg Hold Front
120. Long Jump Jump
121. Long Jump Run
122. Luge
123. Paragliding
124. Petanque Throw
125. Pole Vault Jump
126. Pole Vault Run
127. Pull Ups
128. Punching Sandbag
129. Punching Speed Bag
130. Push Up
131. Quadruped Hip-Extension
132. Racewalk Walk
133. Ride Bike
134. Ride Horse
135. Ride Motercycle
136. Ride Scooter
137. Ride Unicycle
138. Roller Skating Backward

139. Roller Skating Forward
140. Rowing Boat
141. Running In Place Run
142. Scuba Diving
143. Shotput Throw
144. Side Lunge
145. Sit Up
146. Skateboard Forward
147. Skateboard Grind
148. Skateboard Jump
149. Skeleton
150. Ski Backflip
151. Ski Cork
152. Ski Frontflip
153. Ski Jump Land
154. Ski Jump Mid-Air
155. Ski Jump Slide
156. Skydiving
157. Snorkeling
158. Snowboard Jump
159. Snowboard Slide
160. Snowboarding Forward
161. Soccer Dribble
162. Soccer Header
163. Soccer Save
164. Soccer Shoot
165. Soccer Throw
166. Softball Pitch
167. Speedskating Forward
168. Split Leap
169. Sprint Kneel
170. Sprint Run
171. Sprint Start
172. Star Jumping Jump
173. Surfing
174. Swimming Backstroke
175. Swimming Breast Stroke
176. Swimming Butterfly Stroke
177. Swimming Freestyle
178. Taekwondo High Block
179. Taekwondo Kick
180. Taekwondo Low Block
181. Taekwondo Middle Block
182. Taekwondo Punch
183. Tennis Backhand
184. Tennis Forehand
185. Tennis Serve
186. Tire Pull
187. Tire Sled
188. Trapeze Interacting
189. Trapeze Single
190. Triple Jump Jump
191. Triple Jump Run
192. Uneven Bar Cross
193. Uneven Bar Flip
194. Uneven Bar Jump
195. Uneven Bar Land
196. Uneven Bar Spin
197. Volleyball Overhand
198. Volleyball Pass
199. Volleyball Set
200. Volleyball Underhand
201. Water Skiing
202. Weight Lifting Hang
203. Weight Lifting Overhead
204. Weight Lifting Stand
205. Windsurfing
206. Workout Chest-Pull
207. Workout Crunch
208. Yoga Bridge
209. Yoga Cat
210. Yoga Firefly
211. Yoga Tree
212. Yoga Updog

Daily Actions

213. Add New Car Tire
214. Adjusting Glasses
215. ALS Icebucket Challenge
216. Answering Questions
217. Applauding
218. Applying Cream
219. Arm Wave
220. Bandaging
221. Bending Back
222. Blowdrying Hair
223. Blowing Balloon
224. Blowing Glass
225. Blowing Gum
226. Blowing Kisses
227. Blowing Leaf
228. Blowing Nose
229. Bowing Fullbody
230. Bowing Waist
231. Brushing Floor
232. Brushing Hair
233. Brushing Teeth
234. Burping
235. Calfrope Catch
236. Calfrope Rope
237. Calfrope Subdue
238. Carrying With Head
239. Cartwheeling
240. Cast Net
241. Chainsaw Log
242. Chainsaw Tree
243. Chalkboard
244. Chewing Gum
245. Chopping Meat
246. Chopping Wood
247. Cleaning Mirror
248. Cleaning Mopping
249. Cleaning Sweeping
250. Cleaning Vacumming
251. Cleaning Windows
252. Clear Snow Off Car
253. Climb Ladder
254. Climb Stair
255. Climbing Tree
256. Closing Door
257. CPR
258. Crawling
259. Cross Body Shoulder Stretch
260. Cutting Onion

261. Dabbing
262. Dog Highfive
263. Dog Walking
264. Drinking With Cup
265. Drinking With Straw
266. Eat Apple
267. Eat Burger
268. Eat Spagetti
269. Eating Hotdogs
270. Eating Ice Cream
271. Eating Oyster
272. Face Slapping
273. Falling Off Chair
274. Fire Extinguisher
275. Fist Bump
276. Flamethrower
277. Folding Blanket
278. Folding Clothes
279. Gas Pumping To Car
280. Guitar Smashing
281. Hailing Taxi
282. Haircut Scissor
283. Hammering Nail
284. Hand In Hand
285. Hand-Drill Firemaking Blow
286. Hand-Drill Firemaking Drill With Bow
287. Hand-Drill Firemaking Drill With Hand
288. Handsaw
289. Handshake Dog
290. Hanging Clothes
291. Headbang
292. Heimlich Maneuver
293. High Five
294. Hold Baby
295. Hold Baby With Wrap
296. Hookah
297. Hugging Animal
298. Hugging Human
299. Ironing Clothes
300. Jack Up Car
301. Kick Open Door
302. Kiss
303. Leaf Blowing
304. Milking
305. Neck Side Pull Stretch
306. Opening Door
307. Pancake Flip
308. Peeling Banana
309. Pizza Dough Toss
310. Plunging Toilet
311. Pottery Wheel
312. Pouring Wine
313. Push Car
314. Push Wheelchair
315. Push Wheelchair Alone
316. Putting Scarf On
317. Read Newspaper
318. Reading Book
319. Remove Car Tire
320. Rescue Breathing
321. Riding Camel
322. Riding Elephant
323. Riding Mechanical Bull
324. Riding Mule
325. Riding Ostrich
326. Riding Zebra
327. Rolling Snow
328. Salute
329. Screw Car Tire
330. Setup Tent
331. Shake Cocktail
332. Shaking Head
333. Shaving Beard
334. Shoe Shining
335. Shoveling Snow
336. Sledgehammer Strike Down
337. Smoking Exhale
338. Smoking Inhale
339. Spitting On Face
340. Spraying Wall
341. Sticking Tongue Out
342. Stomping Grapes
343. Styling Hair
344. Swinging Axe On A Tree
345. Talking Megaphone
346. Talking On Phone
347. Throwing Bouquet
348. Using Inhaler
349. Using Lawn Mower
350. Using Lawn Mower Riding Type
351. Using Metal Detector
352. Using Scythe
353. Using Spinning Wheel
354. Using String Trimmer
355. Using Typewriter
356. Walking With Crutches
357. Walking With Walker
358. Wall Paint Brush
359. Wall Paint Roller
360. Washing Clothes
361. Washing Dishes
362. Watering Plants
363. Wear Face Mask
364. Wear Helmet
365. Whipping
366. Writing On Blackboard
367. Yawning

Musical Instruments

368. Accordian
369. Bagpipes
370. Bangu
371. Banjo
372. Bass Drum
373. Bowsaw
374. Cajon Drum
375. Castanet
376. Cello
377. Clarinet
378. Conga Drum
379. Cornett
380. Cymbals
381. Doublebass
382. Erhu

383. Gong
384. Grandpiano
385. Guitar
386. Handpan
387. Harp
388. Hulusi
389. Jazzdrum
390. Leaf-Flute
391. Lute
392. Maracas
393. Melodic
394. Noseflute
395. Ocarina
396. Otamatone
397. Panpipe
398. Piccolo
399. Recorder
400. Sanxian
401. Saxophone
402. Serpeng
403. Sheng
404. Sitar
405. Snare Drum
406. Sunoa
407. Taiko Drum
408. Tambourine
409. Thereminvox
410. Timpani
411. Triangle
412. Trombone
413. Trumpet
414. Ukulele
415. Viola
416. Violin
417. Xylophone
418. Yangqin

Games and Hobbies

419. Air Drumming
420. Air Guitar
421. Air Hockey
422. Alligator Wrestling
423. Archaeological Excavation
424. Arm Wrestling
425. Atlatl Throw
426. Axe Throwing
427. Balloon Animal
428. Beer Pong Throw
429. Belly Dancing
430. Blow Gun
431. Building Snowman
432. Card Throw
433. Conducting
434. Decorating Snowman
435. Dice Shuffle Reveal
436. Dice Stack Shuffle
437. DJ
438. Draw Handgun
439. Face-Changing Opera
440. Fire Breathing
441. Fire Dancing Circulating
442. Fish-Hunting Hold
443. Fish-Hunting Pull
444. Flipping Bottle
445. Floss Dance
446. Flying Kite
447. Ganggangsullae
448. Gangnam Style Dance
449. Grass Skating
450. Guitar Flip
451. Hopscotch Pickup
452. Hopscotch Skip
453. Hopscotch Spin
454. Ice Scuplting
455. Juggling Balls
456. Kick Jianzi
457. Knitting
458. Marble Scuplting
459. Moonwalk
460. Piggyback Ride
461. Play Diabolo
462. Play Kendama
463. Play Yoyo
464. Playing Nunchucks
465. Playing Rubiks Cube
466. Playing Seesaw
467. Playing Swing
468. Rock Balancing
469. Rock Paper Scissors
470. Running On Four
471. Sack Race
472. Sand Scuplting
473. Segway
474. Shoot Dance
475. Shooting Handgun
476. Shooting Shotgun
477. Shuffle Dance
478. Sling
479. Slingshot
480. Snow Angel
481. Speed Stack
482. Spinning Basketball
483. Spinning Book
484. Spinning Plate
485. Stone Skipping
486. Sword Swallowing
487. Taichi Fan
488. Taking Photo Camera
489. Taking Selfie
490. Tap Dancing
491. Three Legged Race
492. Throw Boomerang
493. Throw Paper-Plane
494. Tight-Rope Walking
495. Trampoline
496. Tug Of War
497. Underarm Turn
498. Walking On Stilts
499. Whistle One Hand
500. Whistle Two Hands

4. Composite Classes

We list how *Musical Instrument* and *Sports/Athletics* classes form to become composite actions. We list indices of the classes for each composite action.

4.1. Sports/Athletics

1. 49
2. 79, 80
3. 99
4. 65, 66, 67
5. 2
6. 178, 179, 180, 181, 182
7. 120, 121
8. 39, 40, 41, 42
9. 114
10. 140
11. 111, 112
12. 25, 26, 27, 28, 29, 30, 31
13. 60
14. 56, 57, 58
15. 156
16. 144
17. 59
18. 150, 151, 152
19. 168
20. 167
21. 102, 103
22. 145
23. 81, 82, 83
24. 92
25. 128, 129
26. 50, 51
27. 53, 54
28. 138, 139
29. 43
30. 174, 175, 176, 177
31. 183, 184, 185
32. 201
33. 8, 9, 10, 11, 12
34. 142
35. 149
36. 1
37. 32
38. 62, 63, 64
39. 141
40. 109
41. 104
42. 122
43. 110
44. 38
45. 100
46. 157
47. 37
48. 197, 198, 199, 200
49. 116
50. 153, 154, 155
51. 84
52. 131
53. 127
54. 18, 19, 20, 21, 22, 23, 24
55. 117, 118, 119
56. 186, 187
57. 160
58. 169, 170, 171
59. 158, 159
60. 206, 207
61. 13
62. 172
63. 133, 134, 135, 136, 137
64. 123
65. 124
66. 205
67. 5, 6, 7
68. 86
69. 208, 209, 210, 211, 212
70. 113
71. 202, 203, 204
72. 166
73. 105, 106, 107, 108
74. 192, 193, 194, 195, 196
75. 125, 126
76. 61
77. 173
78. 143
79. 85
80. 188, 189
81. 130
82. 101
83. 55
84. 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78
85. 52
86. 115
87. 91
88. 146, 147, 148
89. 87, 88
90. 44, 45
91. 89, 90
92. 3
93. 190, 191
94. 4
95. 35, 36
96. 34
97. 33
98. 14, 15, 16, 17, 46, 47, 48
99. 93, 94, 95, 96, 97, 98
100. 132
101. 161, 162, 163, 164, 165

4.2. Musical Instruments

1. 369, 377, 379, 388, 390, 394, 395, 397, 398, 401, 402, 403, 406, 412, 413,399
2. 371, 373, 376, 381, 382, 385, 387, 391, 396, 400, 404, 409, 414, 415, 416
3. 370, 372, 374, 375, 378, 380, 383, 386, 389, 392, 405, 407, 408, 410, 411, 417, 418
4. 368, 384, 393

5. HAA-COCO

Here we list the classes in HAA-COCO.

- 1, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 32, 33, 34, 35, 36, 37, 45, 46, 47, 58, 60, 80, 81, 82, 86, 87, 88, 89, 99, 108, 110, 111, 115, 124, 125, 127, 128, 132, 133, 134, 135, 136, 139, 142, 160, 161, 162, 163, 164, 165, 182, 183, 184, 196, 197, 198, 199, 201, 202, 203, 212, 214, 235, 236, 237, 245, 246, 248, 249, 250, 251, 252, 263, 264, 265, 266, 267, 268, 276, 277, 278, 289, 298, 299, 305, 307, 311, 312, 313, 314, 316, 317, 318, 321, 325, 328, 330, 336, 337, 339, 345, 357, 358, 359, 360, 361, 367, 375, 376, 379, 381, 383, 384, 386, 388, 398, 400, 409, 410, 411, 412, 413, 414, 415, 427, 431, 434, 435, 443, 454, 480, 481, 487, 488

6. Sample Videos

Figure 1 shows the first frame of a video in different classes. Figure 2 lists diverse videos per class.

7. Hierarchy

Figure 4 shows the hierarchy of action classes in *Sports/Athletics* area where the actions are grouped together with other actions in the same sports category.