Expanding AI Audits to Include Instruments:

Accountability, Measurements, and Data in Motion Capture Technology

Dora Eskridge
School of Data Science
University of Virginia
Charlottesville, VA
de2br@virginia.edu

Iman Yousfi
School of Data Science
University of Virginia
Charlottesville, VA
iy5sw@virginia.edu

Sivaranjani Kandasami
School of Data Science
University of Virginia
Charlottesville, VA
nyc2xu@virginia.edu

Suraj Kunthu
School of Data Science
University of Virginia
Charlottesville, VA
sk9km@virginia.edu

***Abstract*—"Expanding AI Audits" proposes to develop and extend AI audit frameworks to include hardware and other instruments used to collect data used in AI and other data-driven algorithmic applications. The project extends AI audit frameworks to assess the outcomes of an AI system and examine the assumptions on which those systems are based. Doing so enables auditors to assess the validity of an AI system, its appropriateness for use in specific contexts, and the conditions under which such assumptions may fail to produce safe and effective outcomes. The project will examine the use of motion capture technology as a mechanism for data collection and AI development and produce an expanded audit framework for hardware and other mechanisms. This framework will be accompanied by a workshop convening technologists, audit professionals, and regulators to spread the framework and findings and a series of public events.**

***Index Terms*—Audit, Artificial Intelligence, AI, Motion Capture**

I. Introduction

In the advent of 5G and IoT, we are seeing an unprecedented use of Motion Classification Computer Vision AI models. It is used is a variety of fields from intelligence surveillance to human-computer interaction. Motion capture is a complex process and requires many shots of subjects performing various actions. Motion capture experts must then manually extract the data and classify each action and when each action starts and ends. With the onset of AI, there is an increasing number of companies leveraging AI models to automate motion capture classification. However, AI motion capture classification models need to be audited.

Looking back at the history of photography, the calibration performed shows that the baseline chemistry used by photo labs to assess skin tones, shadows, and light during the printing process used only white models for over fifty years. This had a negative impact on the accurate appearance of people of color in photography. Current motion capture classification models are no different and follow a similar bias. Most are trained only on the "Normative" subject: male, white, 'able-bodied,' and of unremarkable weight. What about subjects that do not fit those descriptors? The goal of this project is to audit Computer Vision Action Classification/Pose Estimation Models against our Human-Validated Classification of Movements to determine vulnerabilities such as biases. Our main task is to manually classify actions and their start and end times for 36 diverse participants based off a set of actions given in a well-defined script to prepare these observations for study.

II. Success Criteria

We have two primary deliverables for our project:

1) Strategy
    a) Find a technique to differentiate between motions in video
2) Motion Dataset
    a) Use the technique to complete segregating the existing videos and create a csv with the timestamp breakup

III. Data Summary

The data collected for our project was part of a study performed at NYU Tandon and is protected under the Institutional Review Board (IRB). Therefore, we have a few unique privacy limitations associated with our data. The data set consists of Motion Capture Export from 36 sessions.

- Data Set Size: Approximately 22.06 GB
- Rows/Columns: Binary FBX as well as MP4 and MKV video files
- Nature of Data: Motion capture data in .fbx format, including time-series information
- Data Files:
    - 36 sessions of .fbx files
    - Blurred video recordings of each session (255 MB Screen capture + 245 MB color video side + 30 MB Infrared video left +30 MB Infrared video left) * 36 Sessions * 0.001 = 20.16 GB of total data.

IV. Data Assumptions and Limitations

Motion capture data files are in Autodesk FBX binary format and can be opened with free tools as well as converted into comma separated value format (CSV).

Due to data protection under the IRB, we cannot download any of the data files locally. We will be using only virtual machines for the project. Additionally, we cannot share specific data and information outside of the study team.

In terms of technological limitations, the data format will prove challenging. For each participant's motion capture session, we have the raw data in three forms: video of the session, csv files of the movements along an xyz plane, and the fbx files capturing the motion capture data. The video will provide us with a strong understanding of how the participant is moving at a given moment. However, the csv does not indicate which body part is being tracked in each column. Rather, each individual body part has the same identifying name as “bone”. Additionally, we will not be able to read or use the fbx data without additional processing through AutoDesk motion capture software, a platform that we have not used previously.

V. Summary of Data Processing, Data Aggregation

The original study and data capture was completed in the Media Lab at NYU Tandon with a cross-site collaborative research team. The team collected the data in two separate sessions. For the data collection, each participant wore a motion capture suit to accurately capture their body’s movements. Once the participant was prepared for data collection, the team read aloud the movements script, ensuring that each participant completed the same motions in the same order.

The data was collected in two batches. 24 participants were processed in June, and 24 participants were processed in July, with 12 individuals participating in both data collections. The raw data provided from these sessions includes a csv file documenting the xyz plane positioning of each participant’s relevant body parts throughout the recording, an .fbx file containing the motion capture data, raw video footage of the session, and the script that was read to the participants.

In the context of our task, the target variable is the video’s time stamps. Specifically, we need to segment the videos into distinct movements based on our selected technique. Potential predictor variables will principally be drawn from the csv file, as movements along the xyz plane will be the driving factor in determining when movements begin and end.

"In order to extend embedding techniques to entire datasets, we have to define a distance between datasets,"

"generalizes the notion of the shortest path between two points to the shortest set of paths between distributions"

VI. Data Visualizations

The image below displays the live motion capture and the wireframe interpolation on AutoDesk Motion Builder. The red box surrounding the software application window indicates the session is being recorded. After a session is complete, that motion study can be exported as a .csv file.

VII. Computer Vision Literature Review

A. Human Pose Estimation:

Our team engaged with the examination of the current state of Motion Capture Technology from an ethical, inclusive and technical perspective. There have been many great advancements in Human Pose Estimation (HPE) using Deep Learning methodology. HPE is a type of computer vision task that identifies points on the human body. Once all connected, it creates skeleton-like representation of the human body. HPE is typically used in gaming, movies, augmented reality, and many other places. There are different levels of identifying these points on the human: 2D, 3D and 6D pose estimation. In the literature we found two dominant approaches: CNNs and Transformers with promising results for these types of computer vision tasks. After understanding the computer vision techniques that were being employed for the classification algorithms, we wanted to gauge a better understanding of the datasets they were used to train from an AI Auditing perspective. The results we found from these benchmarking datasets were unfortunately harrowing. The standard HPE benchmark datasets are Human3.6M and MPI-INF-3DHP that are widely used for training, representing solely normative assumptions of the human body. The datasets were predominantly men, white, “able-bodied”, and unremarkable weight. There is a clear lack of diversity, which gets encoded into these computer vision models that are to use classify and make decisions.

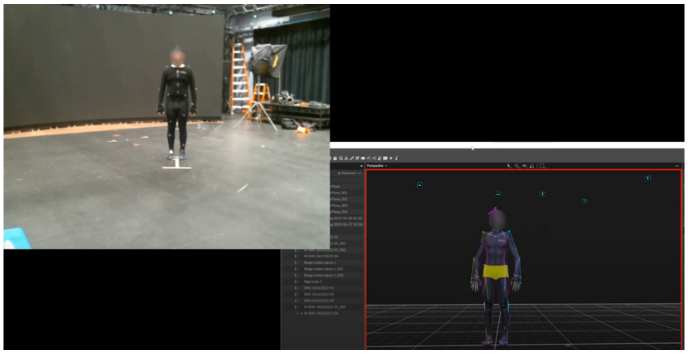
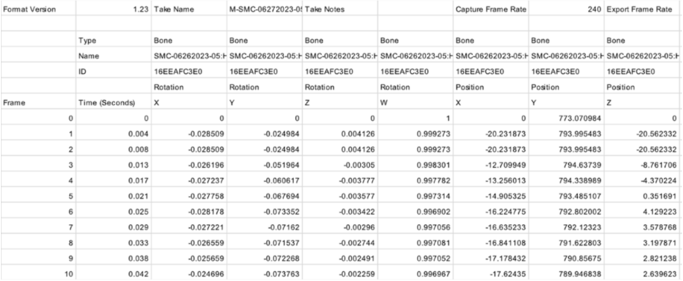


Fig. 1: Screenshot of Motion Capture Session



Format Version	1.23	Take Name	M-SMC-06272023-0!	Take Notes		Capture Frame Rate	240	Export Frame Rate
	Type	Bone	Bone	Bone	Bone	Bone	Bone	Bone
	Name	SMC-06262023-05:H	SMC-06262023-05:H	SMC-06262023-05:H	SMC-06262023-05:H	SMC-06262023-05:H	SMC-06262023-05:H	SMC-06262023-05:H
	ID	16EEAFC3E0	16EEAFC3E0	16EEAFC3E0	16EEAFC3E0	16EEAFC3E0	16EEAFC3E0	16EEAFC3E0
		Rotation	Rotation	Rotation	Rotation	Position	Position	Position
Frame	Time (Seconds)	X	Y	Z	W	X	Y	Z
0	0	0	0	0	1	0	773.070984	0
1	0.004	-0.028509	-0.024984	0.004126	0.999273	-20.231873	793.995483	-20.562332
2	0.008	-0.028509	-0.024984	0.004126	0.999273	-20.231873	793.995483	-20.562332
3	0.013	-0.026196	-0.051964	-0.00305	0.998301	-12.709949	794.63739	-8.761706
4	0.017	-0.027237	-0.060617	-0.003777	0.997782	-13.256013	794.338989	-4.370224
5	0.021	-0.027758	-0.067694	-0.003577	0.997314	-14.905325	793.485107	0.351691
6	0.025	-0.028178	-0.073352	-0.003422	0.996902	-16.224775	792.802002	4.129223
7	0.029	-0.027221	-0.07162	-0.00296	0.997056	-16.635233	792.12323	3.578768
8	0.033	-0.026559	-0.071537	-0.002744	0.997081	-16.841108	791.622803	3.197871
9	0.038	-0.025659	-0.072268	-0.002491	0.997052	-17.178432	790.85675	2.821238
10	0.042	-0.024696	-0.073763	-0.002259	0.996967	-17.62435	789.946838	2.639623

Fig. 2: Screenshot of Exported Motion Capture Study

B. Activity Recognition:

Upon finding this, we transitioned our technical literature review to better understand these benchmark datasets. Our findings led us to better understand Computer Vision techniques used in the Human Action Recognition space. The team found that the standard technique used in the literature is Sequential Bag-of-Word (BoW) models. This technique is typically known to be a NLP technique for Document Classification. In Computer Vision, it is applied similarly for image classification. Like words, it treats the images features as words, and creates a vector occurrence of counts of a vocabulary of local image features.

Description automatically generatedFirst, the local features are extracted from the videos. Then, these videos are treated as

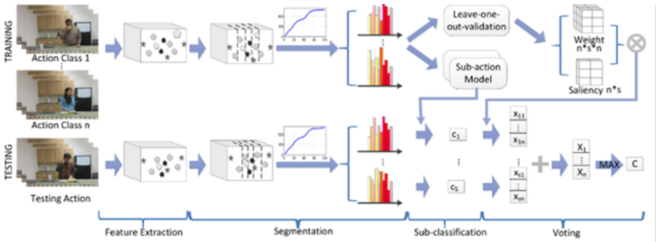


Fig. 3: Activity Recognition Pipeline

building blocks of visual words. Then the video is segmented into smaller clips. The measure the distances between these smaller parts and split them into equal sections, which represent the different parts of the action. Next, we create graphs that show how often different actions happen in order. For example, if someone is running, jumping, then climbing, the graph would show that order. Before we decide what action is happening, we compare these smaller action parts to ones we've seen before (training data). These similarities act like votes for what action is happening. We also look at how accurate our predictions were before and assign importance to different sections based on that accuracy. Finally, we use a voting system to decide what action is happening. We count all the votes for each possible action and choose the one with the most votes as our final guess. [1]

VIII. Technical Literature Review & Analysis

Based on our sponsor's request, our team conducted an extensive analysis within the realm of activity recognition literature. Our investigation encompassed a broad spectrum of datasets and data science techniques, delving into various types of activities and motions. We emphasized scrutinizing pivotal areas concerning activity recognition in video data. We investigated several key questions per our sponsor's guidance—How were the datasets constructed? What activities were defined, and how were they marked? Was there any data cleaning? For a full dataset-specific breakdown of the technical literature review, please see the appendix at the end of this document.

Throughout our exploration, several general findings emerged:

- Timestamps or Activity Tagging for Action Identification: A prevalent trend across the datasets was the consistent utilization of timestamps or activity tagging to pinpoint and categorize different actions within the data. This systematic approach facilitated the dataset's organization and provided a clear reference point for understanding the sequence of activities.
- Unique Identifiers for Non-Action Time Periods: Interestingly, a significant number of datasets featured a unique identifier dedicated to non-action time periods. This inclusion serves a crucial purpose by distinguishing intervals devoid of any notable activity, aiding in the dividing of action boundaries and enhancing the accuracy of activity recognition algorithms.
- Baseline Identification of Actions using Scripts or Actor Narration: A notable observation was the prevalent practice of establishing a baseline for action identification through scripts or actor narration. By providing contextual information or explicit instructions regarding the actions performed, datasets equipped researchers with valuable insights into the intended activities, thereby facilitating more accurate annotation and analysis.

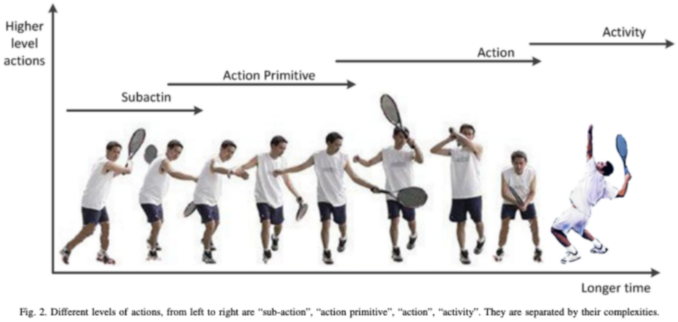

Higher level actions
Activity
Action
Action Primitive
Subactin
Longer time
Fig. 2. Different levels of actions, from left to right are "sub-action", "action primitive", "action", "activity". They are separated by their complexities.

Fig. 4: Activity States [1]

In essence, our deep dive into technical literature revealed several recurring patterns and methodologies employed within the domain of activity recognition. These findings not only shed light on the prevalent practices but also serve as a valuable resource for informing future research endeavors and dataset construction initiatives in the field.

These findings aided us in refining the action-defining technique, enabling us to effectively apply this strategy to the participant data. Much like in some of the articles read in the technical review, we are given the same script that was read out loud to the participants of the study. They provided an excellent baseline, so we know what action to expect and in what order. Below is the list of actions from the provided script.

- Script
  - Start T-Pose
  - Walk
  - Sitting x2
  - Rotate
  - T-Pose
  - Squat
  - Arm Swing
  - Arm Circles
  - Jump x2
  - Kick
  - Carrying Kettlebell in front
  - Carrying Kettlebell in side
  - Ball Carrying
  - Stepping
  - Ball Bouncing
  - Cross Toe Touching x2
  - Running On Spot
  - Phone Usage
  - Pointing
  - Watch Checking
  - Staying Warm
  - Cross Arms

- Plank
- End T-Pose

IX. Methodology: Defining Action

Our team's primary responsibility for this project was to devise a methodology to determine what constitutes a motion in the dataset. The technique also encompasses how to determine the exact start and end time of a given motion. Secondarily, we applied this technique to the dataset. In Liu et. all [1], the authors introduce the concept of action primitive. which they defined as components of actions. Our team drew inspiration from the concept that motion consists of smaller actions. These actions help lead up to the activity. We defined motion into four parts as follows:

- Prior Action Primitive: Actions that lead up to an activity.
- Activity Start Time: The actual activity start time.
- Activity End Time: The actual activity end time.
- Post Action Primitive: The actions that follow the end time of activity.

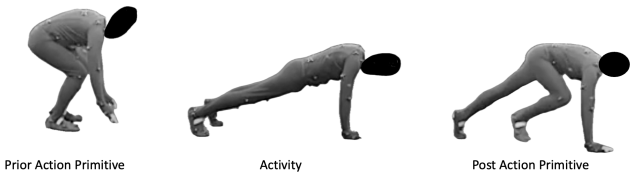


Fig. 5: Motion Definition

The final product for the research team is a csv file of time stamps that indicate motion start and stop times. The file includes the timestamps and the respective definitions for:

- Prior Action Primitive
- Activity Start Time
- Activity End Time
- Post Action Primitive

X. Potential Concerns and Blockers

Based on the literature review, we understand eliminating bias in motion capture models is important. But given the limited dataset and timelines of the capstone, addressing bias may have to be considered out of scope for our project.

XI. Findings and Deliverables

Through our literature review and extensive study of existing motion capture data and techniques, we produced two key deliverables: a decision-making technique for action definition, and a csv of timestamps constructed using that framework. We understand that data fairness, and an awareness of diversity, equity, and inclusion, have to be embedded in a process to ensure that bias does not creep into the system over time, so the inclusion and accurate recognition of diverse body types was a primary focus of our team when establishing our motion-defining framework. With our action definitions, and with the breakdown of prior action and post action primitives, we aimed to create descriptive and inclusive identifiers for each documented action.

The actions were not defined strictly by the movement of the normative body, and they were not expected to fit perfectly into the constraints of normative body movement. The definitions allow participants to move in their own way – for example, consider the plank movement. With a prior action primitive that contains movement from the neutral standing position to the participant's specific plank position, and a post action primitive that contains all movement back to neutral standing position, all body types can be recognized and correctly annotated in this action, regardless of movement speed or ability level.

After establishing an inclusive framework, we created the csv of timestamps for all current video data in the study, effectively annotating the participant data with a decision-making technique that eliminates any subjectivity and opportunity for bias .

XII. Future Work

Moving forward, the research team has received our deliverables, and they have the detailed breakdown of the action-definition technique for future use. They will continue to expand AI motion capture auditing, testing their levels of accuracy against off the shelf models and identifying any biases in the current systems. Through their work, they will drive increased inclusivity and accountability in the world of motion capture data.

Appendix

A. Dataset 1: Breakfast: A Large-Scale Database for Video Understanding

Relevant URLs: breakfast-actions-dataset, paper_cameraReady-2

How was the dataset constructed: 10 actions were performed by 52 individuals in 18 kitchens, with between 3-5 cameras per session. The data is meant to capture realistic natural action rather than lab environment movements. The videos are normalized to consistent FPS and pixels, the actors were not scripted or directed, but rather just handed a recipe and told to prepare the listed food items.

Activities defined: There were 10 activities total– the preparation of:

1) coffee
2) orange juice
3) chocolate milk
4) tea
5) cereal
6) fried eggs
7) pancakes
8) fruit salad
9) sandwich
10) scrambled eggs

How are activities marked: Activities are annotated as coarse actions and "silence" samples. Coarse actions are made up of a series of finer actions that often take place in different orders for different participants. "Silence" samples are the moments in between coarse actions.

Any data cleaning: No.

Other relevant notes: They had an interesting idea to use speech recognition techniques applied to video data for activity recognition.

B. Dataset 2: The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines

Relevant URLs: https://epic-kitchens.github.io/2018, https://arxiv.org/pdf/1804.02748.pdf

How was the dataset constructed: This dataset contains first-person headcam recordings of non-scripted actors in their own kitchens. The videos were annotated using live audio feed narration. Videos were filmed across 32 kitchens in 4 cities, and actors wore the headcam whenever they were in the kitchen. The full dataset contained over 50 hours of recordings, and each recording was annotated with start and end times for each action, as well as lined boxes around objects subject to user interaction.

Activities defined: There were almost 40,000 action segments, with a massive number of actions defined. The dataset also defines objects and object interactions in addition to actions. The dataset defines actions with combinations of verb and noun classes.

How are activities marked: Activities are marked based on the audio commentary. The actor will narrate their action ("I am opening miso paste") and the action is identified based on the key words given (open, paste).

Any data cleaning: Not discussed, but it seems the researchers kept all video data and annotated relevant clips.

Other relevant notes: Actions were annotated based on the narration of the actor, as they are the expert on what they are doing, but were quality checked by a team of annotators and then confirmed via random sampling and manual viewing by the research team.

C. Dataset 3: ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding

Relevant URLs: https://ieeexplore.ieee.org/document/7298698

How was the dataset constructed: They heavily relied on the crowd and specifically, Amazon Mechanical Turk, to help acquire and annotate ActivityNet.

Activities defined: ActivityNet provides samples from 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 video hours.

How are activities marked: The videos are manually trimmed.

Any data cleaning: The videos were trimmed into sub-actions.

D. Dataset 4: UCF101

Relevant URLs: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/UCF101_CRCV-TR-12-01.pdf

How was the dataset constructed: This benchmark dataset was collected from 13,320 YouTube videos.

Activities defined: There are 101 action categories defined from 13320 videos. It is the largest in terms of diversity of actions. The action categories for UCF101 data set are: Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Baseball Pitch, Basketball Shooting, Basketball Dunk, Bench Press, Biking, Billiards Shot, Blow Dry Hair, Blowing Candles, Body Weight Squats, Bowling, Boxing Punching Bag, Boxing Speed Bag, Breaststroke, Brushing Teeth, Clean and Jerk, Cliff Diving, Cricket Bowling, Cricket Shot, Cutting In Kitchen, Diving, Drumming, Fencing, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Golf Swing, Haircut, Hammer Throw, Hammering, Handstand Pushups, Handstand Walking, Head Massage, High Jump, Horse Race, Horse Riding, Hula Hoop, Ice Dancing, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Knitting, Long Jump, Lunges, Military Parade, Mixing Batter, Mopping Floor, Nun chucks, Parallel Bars, Pizza Tossing, Playing Guitar, Playing Piano, Playing Tabla, Playing Violin, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Sitar, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rafting, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Shaving Beard, Shotput, Skate Boarding, Skiing, Skijet, Sky Diving, Soccer Juggling, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Swing, Table Tennis Shot, Tai Chi, Tennis Swing, Throw Discus, Trampoline Jumping, Typing, Uneven Bars, Volleyball Spiking, Walking with a dog, Wall Pushups, Writing On Board, Yo Yo.

How are activities marked: They use a Sequential Bag of Words technique

Any data cleaning: The videos are downloaded from YouTube and the irrelevant ones are manually removed. All clips have fixed frame rate and resolution of 25 FPS and 320 × 240 respectively. The videos are saved in .avi files compressed using DivX codec available in k-lite package. The audio is preserved for the clips of the new 51 actions.3

Other relevant notes: Thought this is the most diverse dataset in the literature, it disproportionately highlights activities with physically-able adults.

E. Dataset 5: THUMOS Challenge: Action Recognition with a Large Number of Classes

Relevant URLs: https://www.crcv.ucf.edu/THUMOS14/home.html

How was the dataset constructed: Dataset has action recognition and temporal action detection tasks and contains 4 parts:

- Training Data
- Validation Data
- Background data
- Test Data

The training data comprises the entire UCF101 action dataset for action recognition, encompassing 101 human action categories with 13,320 temporally trimmed videos, while for temporal action detection, a subset of the UCF101 action dataset with 20 action classes is utilized, with the option to incorporate the remaining action classes if necessary.

In the validation phase, 1,000 videos are allocated for action recognition, each action class having precisely 10 videos, accompanied by video-level label information indicating primary and secondary actions, albeit all videos remain temporally untrimmed. For temporal action detection, 200 videos from 20 action classes serve as validation data, featuring temporal annotations detailing the start and end times of action instances.

Background data, totaling 2,500 videos exclude instances of any of the 101 or 20 action classes, depending on the task, with each background video pertinent to an action class and the primary class for reference.

Test data has 1,574 temporally untrimmed videos for both action and temporal recognition, which may include instances of one or multiple action classes or none at all.

Activities defined: 101 human action categories (https://www.crcv.ucf.edu/THUMOS14/Class%20Index.txt)

How are activities marked: Activities are marked based on the presence or absence of specific action classes in test videos, alongside a confidence score.

For the Action Recognition task, each test video's prediction involves outputting a real-valued score indicating the confidence of the predicted presence of the action class within the video. In contrast, for the

Temporal Action Detection task, the prediction not only involves the presence of the action class but also its temporal localization, i.e., the starting and ending times of each detected instance. Ground truth annotations for the temporal locations of action instances are provided for validation videos evaluation.

Any data cleaning: No reference of data cleaning.

Other relevant notes: Defining evaluation metric for the tasks for precise task recognition.

F. Dataset 6: Activity Recognition in the Home Using Simple and Ubiquitous Sensors

Relevant URLs: https://courses.media.mit.edu/2004fall/mas622j/04.projects/home/TapiaIntilleLarson04.pdf

How was the dataset constructed: Data collection methods includes video recordings of tutoring sessions, audio recordings, transcripts of verbal interactions, observational notes, surveys, and questionnaires to capture both qualitative and quantitative data about the tutoring context, learner behavior, and tutor behavior.

Activities defined: Tutoring Activities and Learning Activities.

How are activities marked: Bayesian classifiers were used to detect activities using the tape-on sensor system.

Any data cleaning: No reference of data cleaning.

Other relevant notes: Activity recognition system architecture: The proposed system consists of three major components: (1) The environmental state-change sensors used to collect information about use of objects in the environment, (2) the context-aware experience sampling tool (ESM) used by the end user to label his or her own activities, and (3) the pattern recognition and classification algorithms for recognizing activities after constructing a model based on a training set.

G. Dataset 7: Charades: A Large-Scale Dataset for Video Understanding

Relevant URLs: https://arxiv.org/abs/1804.09626

How was the dataset constructed: Researchers recruited crowd workers over the internet to self-record 1st person and 3rd person actions with a given script. Data has an 80/20 split, with 8.72 activities per video on average.

Activities defined: Scripts come from the Charades dataset.

How are activities marked: Script is provided to workers.

Any data cleaning: Not mentioned.

Other relevant notes: Although actions are the same, 1st person and 3rd person actions are picked up differently. For 1st person, users had the option to hold the camera on their forehead or create a head mount. One arm vs Two arms. The latter option had a monetary incentive of a $0.50 bonus.

H. Dataset 8: Moments in Time Dataset: one million videos for event understanding

Relevant URLs: https://arxiv.org/abs/1801.03150

How was the dataset constructed: The dataset contains one million short videos each with a label corresponding to an event unfolding in 3 seconds, which is the average duration of human memory.

Activities defined: Opening various things, closing various things.

How are activities marked: Each video is tagged with one action or activity label among 339 different classes).

Any data cleaning: Background clutter mentioned but no cleaning.

Other relevant notes: Compound activities that occur at longer time scales can be represented by sequences of three second actions (e.g. agent and scene).

References

[1] https://www.sciencedirect.com/science/article/pii/S2468232216300403