EgoK-360

Launching a 360° Video Dataset for Human Egocentric Action Classification

Nadezhda Bzhilyanskaya Mario De La Garza Liane Vásquez-Weber Dr. Yan Yan

Introduction

Our goal was to create a 360°, egocentric, kinetic video dataset. The purpose of this dataset is for action recognition after being trained using a variety of machine learning algorithms, particularly neural networks. Some action recognition algorithms employed recently include: 3D convolution and so-called Two-Stream neural networks, the latter of which have been successfully used to capture both spatial and temporal information from videos.

To the best of our knowledge, our main contribution is this being the first dataset to encompass all three of these features: egocentric, kinetic (for action recognition) and 360°.

We collected data for 49 actions, totaling approximately 11,300 video clips of at most 10 seconds. Out of these clips we trained a model on 5197 of them. To date, the 3D convolution neural network has been run with variations to its hyperparameters while preparations are being made to run Two-Stream.

Background

Kinetic action recognition video datasets as well as egocentric video data and general 360° videos have already to some degree been created. The following are some examples:

Dataset	Year	Actions	Clips per Action	Total Clips	Orig Videos Unclipped
HMDB-51	2011	51	min 102	6,766	3,312
UCF-101	2012	101	min 101	13,320	2,500
ActivityNet-200	2015	200	avg 141	28,108	19,994
Kinetics	2017	400	min 400	306,245	306,245

Above is a list of kinetic datasets which are available. These sets, although used for action recognition, do not claim to be egocentric and none are 360°.

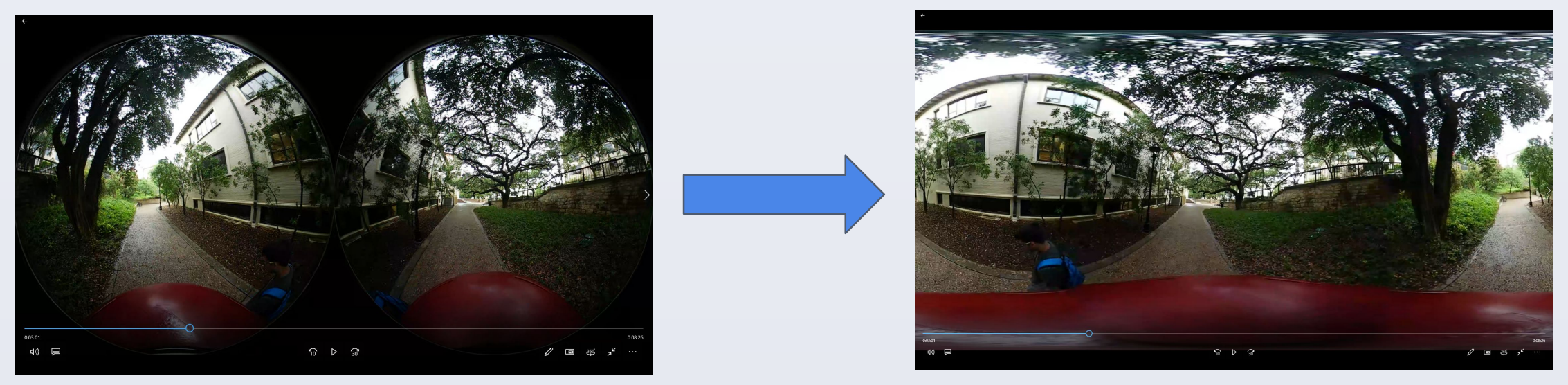
	Context	Hardware
HUJI Ego Seg	Bus, Bike, Horse, Drive, Cook, etc.	GoPro Hero 3+ (& YouTube)
GeorgiaTech First Person Social Interactions	Disney World	Cap Camera
UT Ego	Daily Life	Looxcie Wearable

In the table directly above are all egocentric datasets, however they are not filmed in 360°. These data sets are most similar to ours as many capture normal human actions but ours is unique as it is 360° as well.

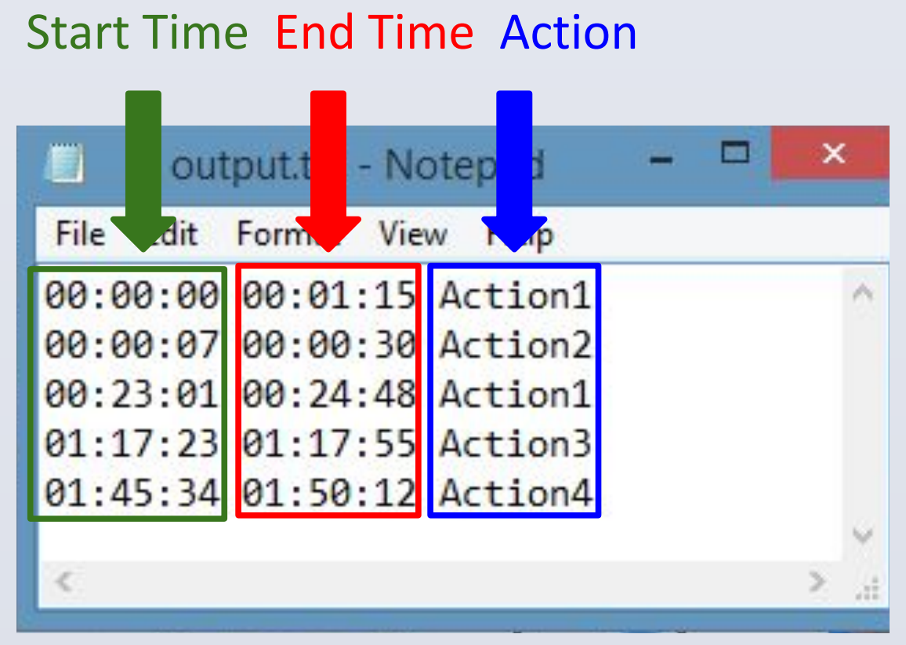
Process

To collect the data we used the Samsung Gear 360° camera, mounted on top of a bicycle helmet. Multiple people wore the camera for the same action to ensure the network would be exposed to different heights and mannerisms. This increases the variety in our dataset—a key element for a dataset when evaluating machine learning algorithms.

Using ActionDirector, the Samsung software for the Gear 360, we created an equirectangular projection of our videos (as shown below).



We annotated our videos to denote when a specific action was occurring and saved the annotations in a text file (as shown below).

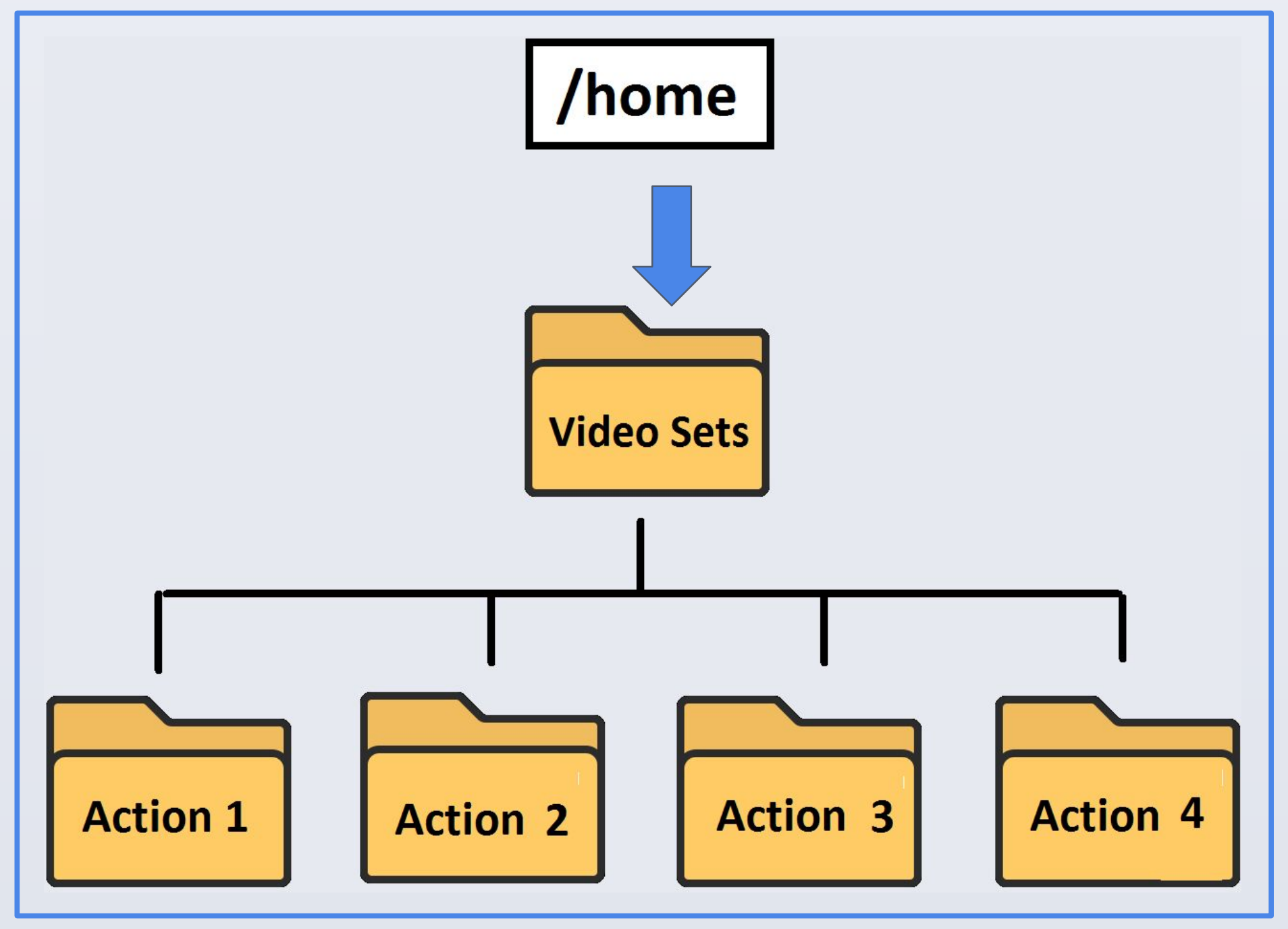


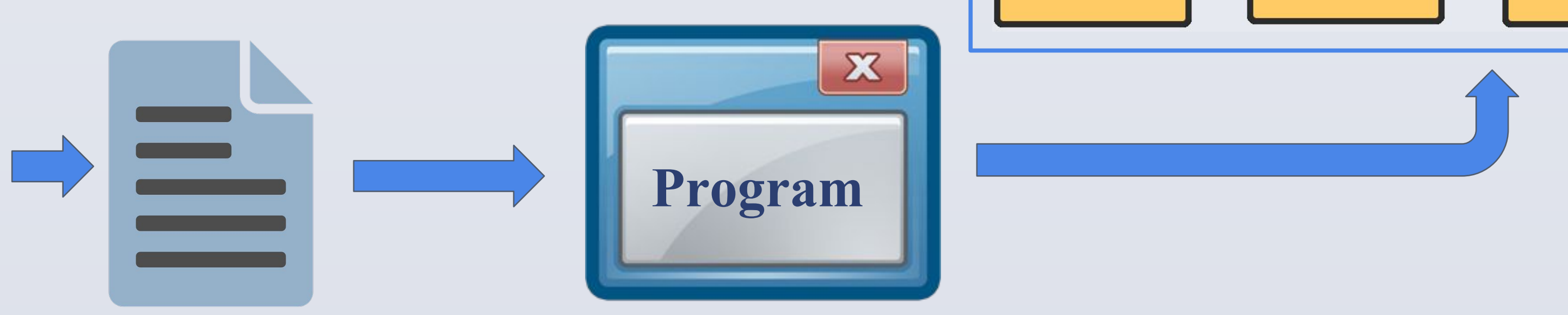
Start Time End Time Action

output.t - Notep d

```
00:00:00 00:01:15 Action1
00:00:07 00:00:30 Action2
00:23:01 00:24:48 Action1
01:17:23 01:17:55 Action3
01:45:34 01:50:12 Action4
```

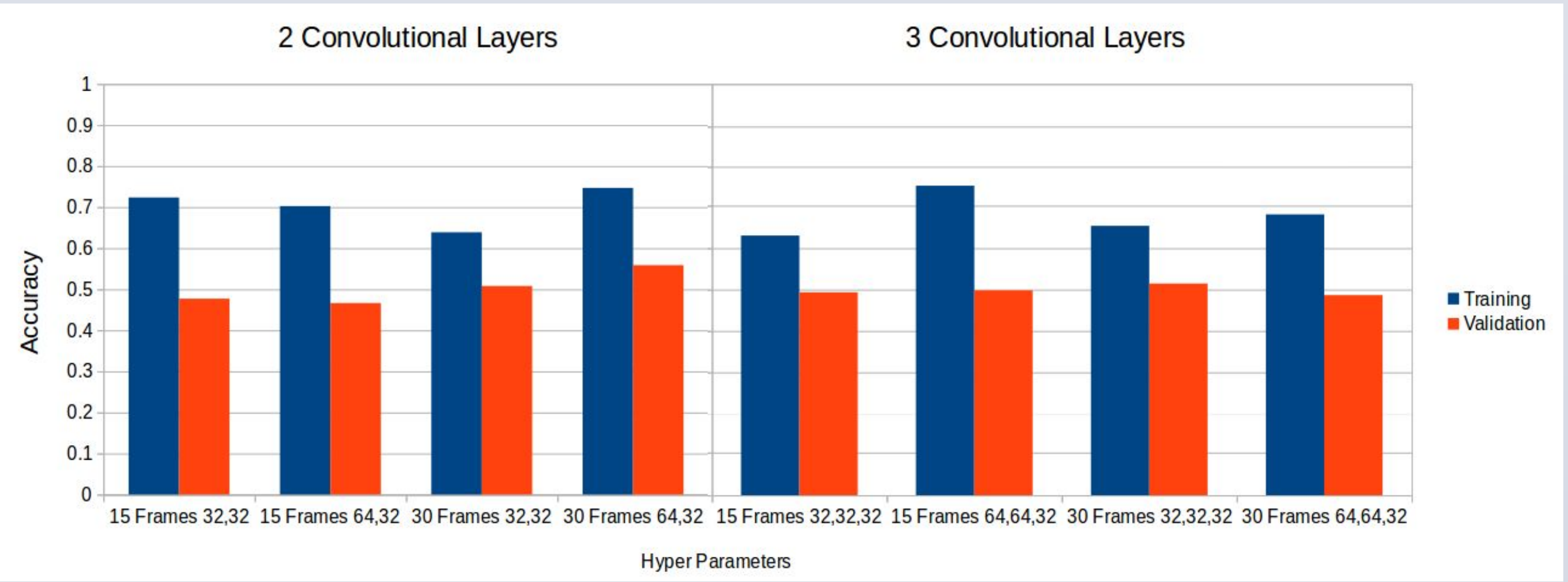
Using a program we wrote, we took the videos and annotation files as input and outputted segments of the videos and stored the dataset in a system of directories.

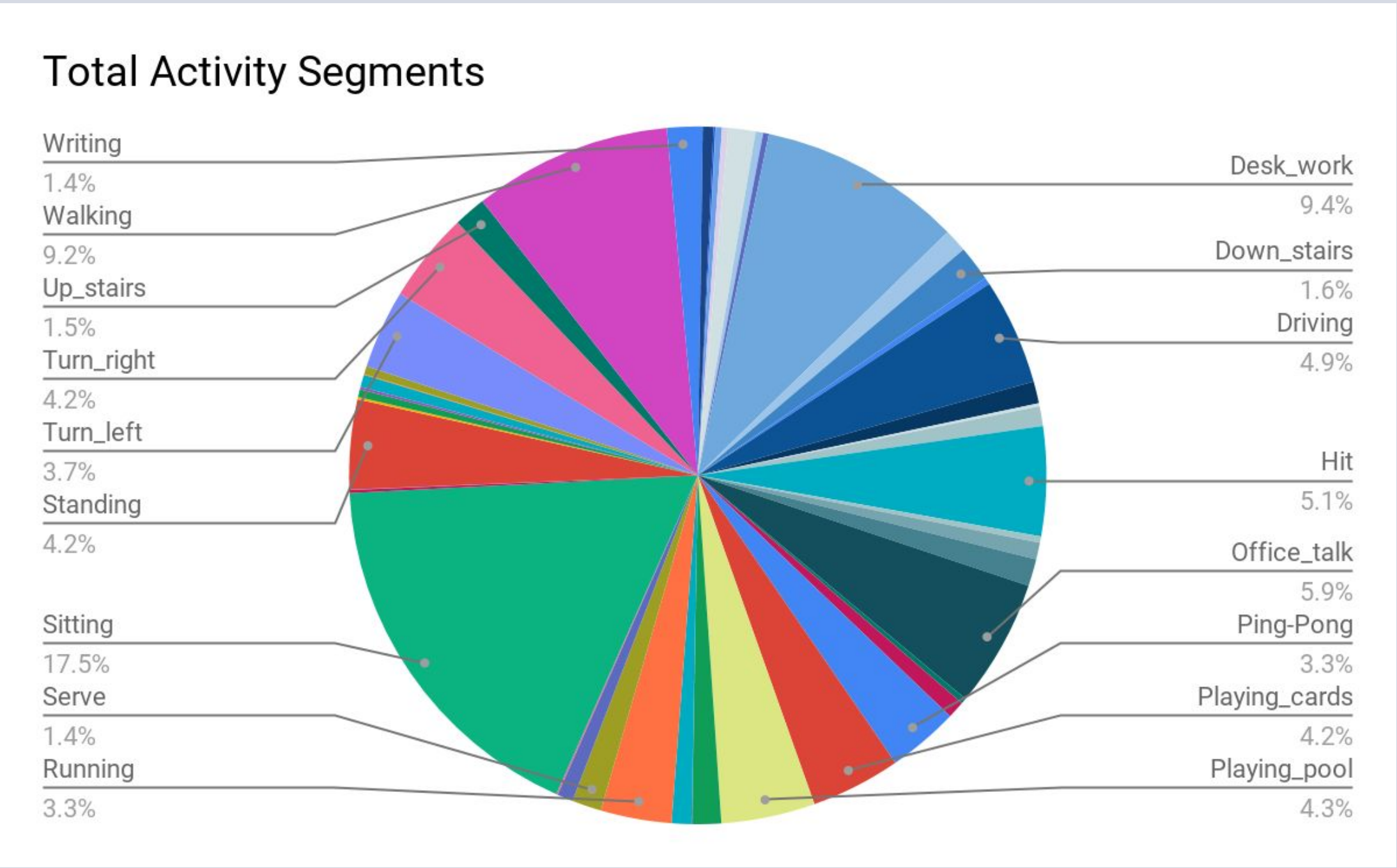




Results

We ran our data through several neural network models using 50 epochs in which we tried using 2 and 3 convolutional layers as well as 15 or 30 frames. All of these attempts yielded similar results and it will take us a lot more testing to try to find a more specialized model that is more accurate.





Dataset	Year	Actions	Clips per Action	Total Clips	Orig Videos Unclipped
EgoK-360	2018	49	avg 232	11,364	127

Our dataset has a wide variability of number of clips per action, from two for rare actions (such as “rack up” in pool) to 1,993 for sitting, which overlaps with many smaller actions. Despite our much smaller number of original videos, we were able to extract many different actions from one video. This is a testament to how activities unfold in real time in life.

Conclusion

We contribute what to the best of our knowledge is the first 360°, Egocentric, Kinetic, video dataset for action classification. We include videos of the same action being repeated in several different settings, thus offering other research teams the needed variety when training their machine learning algorithms.

Additionally, we offer a customized program to automate the video segmentation process.

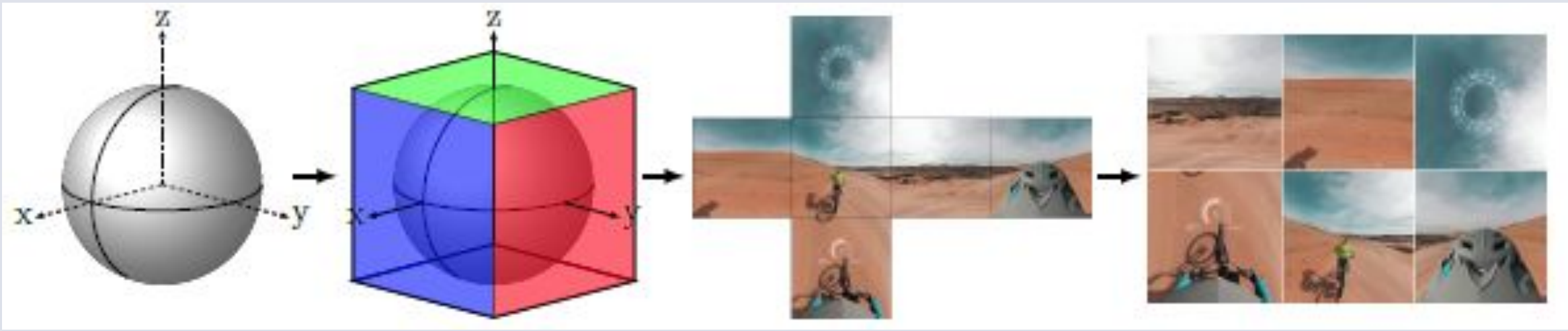
Future Work

This endeavor was a launching of a 360° egocentric kinetic video dataset. With 49 classes and ~ 51% of them containing less than 100 samples, increasing the number of videos in these classes is necessary to compete with available datasets in the context of action recognition.

The current research has the prospect of training various networks, including a two-stream and 3D-conv with LSTM module in addition to the 3D conv network for which results are being obtained

The projection method used was not one that was the least distorting; different projection methods may prove more useful for accurate classification.

Below is an example of the cubemap projection which can increase accuracy (second reference).



References

1) Kay, Will et. al, The Kinetics Human Action Video Dataset, 2017.

2) Su, Yu-Chuan & Grauman, Kristen, Learning Compressible 360° Video Isomers, 2017.

Acknowledgements

We thank the National Science Foundation (NSF) for funding this research under the Research Experiences for Undergraduates Program (CNS-1358939, CNS-1659807) and the infrastructure support provided by the NSF-CRI 1305302 award.

We also thank our mentor, Dr. Yan Yan, and his Ph.D. student, Hao Tang, who have provided our team with guidance and facilitated solutions to problems we would have otherwise spent valuable research time on.