Study of Background Bias in Kinetics-400 Video Action Dataset

Timothy Han
Seoul Scholars International
danny24623@gmail.com

Abstract

Artificial intelligence (A.I.) is helping various parts of human lives. Video action recognition, a field in A.I., is a task of detecting human action given a video. While current state-of-the-art methods show promising results in video action recognition tasks, it is widely believed that many machine learning models rely heavily on the background scene, rather than looking at the human body, to detect the action from a video. While such problem is believed, there has not been a good method to study the problem, as there lacks video dataset where actions were performed under irrelevant backgrounds. To solve this issue, we collected a variety of data, ones depicting an action done in a relevant background, and others in an incorrect, "weird" background. With this dataset, we show that the A.I. relies on a data's background to determine human action, instead of looking at the action itself. [1]

1. Introduction

A.I. contributes many things to our society. It is used everywhere in our daily lives from automatic vacuum cleaners to self-driving cars, and even assists surgeons in a difficult surgery. Even when we use our phones, we use facial recognition to unlock them and when we type, we get our spelling mistakes corrected. Currently, machine learning is the most popular method of artificial intelligence, which to use data to teach the A.I. to do a specific task. Its strength is that it doesn't need an explicit problem solving algorithm, thus giving possibility to solve complex problem, but has the risks of also suffering from the data's issues.

One of machine learning problems include issues in video action recognition. The task of video action recognition is to analyze a video and determine what the human is doing. Suffering from a badly designed dataset, the system can make an error, where instead of focusing on the person in the video, it focuses on the video's background. For example, think of a video where a person is doing a backflip on a golf course. The A.I. model would output *Playing Golf*, instead of *Backflip*, as the A.I. model is focusing more on the background.

The issue above is well known, but hasn't been thoroughly studied. It's because there hasn't been a "weird background" dataset. In this paper we introduce WeridKinetics-300, a manually collected video action dataset. There are 10 classes, a subset of classes in Kinetics-400 [7], including *Tennis*, *Basketball*, *Push Ups*, *Jogging*, *Brushing Teeth*, *Shaving Beard*, *Eating Cake*, *Flipping Pancake*, *Using Computer*, and *Reading Book*. Each class has 30 videos each, where half of the videos are recorded in a relevant background, while the other half is recorded from an irrelevant background. See figure 1 for sample frames of videos.

2. Related Works

2.1. Video Action Recognition

Video action recognition is a task in computer vision where the goal is to identify the action happening in a video. For example, if the video shows a person doing a ballet, the machine learning model needs to output 'Doing a Ballet'. This requires not only understanding the spatial information, e.g., body movement and object location in a given frame, but also temporal understanding, i.e., how the body moves throughout different frames. This makes that task to be more challenging than object recognition task in image that only requires a spatial understanding in a image.

Researchers have been working on video action recognition for a long time. Recently, with advent of deep learning and it's superior performance in object recognition [8], researchers tried using same idea that were used in object recognition on video action recognition as well. Naturally, naïve method that researchers used was to run object recognition model on videos, on each single frame, and average the final output. However in this case, the model is trained to output the action class, instead of object class. TSN [15] is one example, which surprisingly shows good performance of 91.8% accuracy in UCF101 [14] dataset. Recently, there

[1] Please check our project website https://timothydhhan.github.io/weirdkinetics to access the dataset.

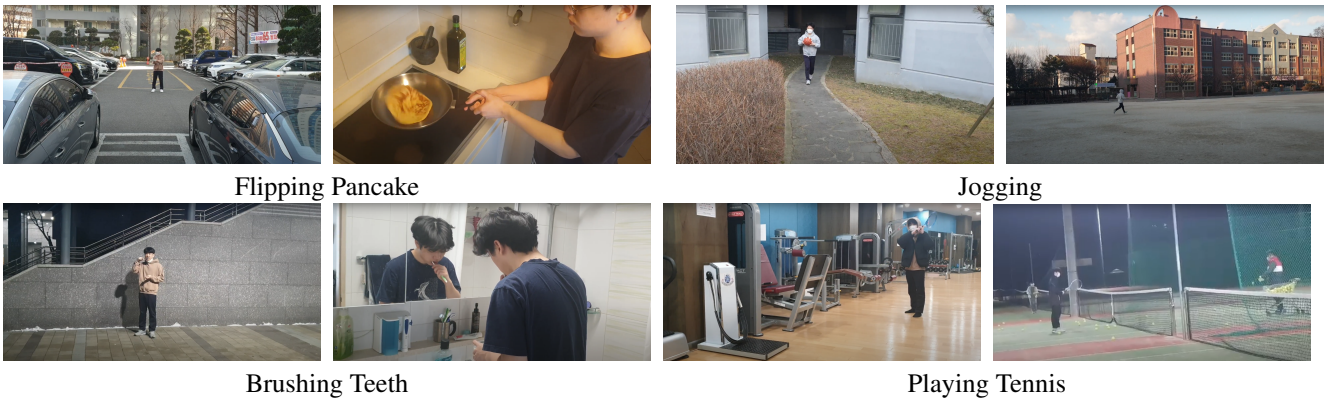

Figure 1. Two samples per action: The left shows videos with unrelated background, where the machine learning model has confused, while the right shows videos with related background, where the machine learning model has guessed correctly.

has been better machine learning designs [2, 3, 13] that understand the temporal information better. I3D [3], for example, considers a video as an 3D-image, and tries to understand temporal information in a same method as the spatial information. In this paper we use I3D [3] model as our main video action recognition tool, given it's popularity among the field.

Concurrently, there has been abundant release of video action datasets with improving quality over time. Video action datasets contains multiple videos on fixed action classes. These fixed action classes serve as a guideline for action recognition models. For example, Kinetics-400 [7] contains 400 action classes with around 500 videos per class. Video action recognition model trained on Kinetics-400 will only output one of the names in 400 action classes when a video is given. In this paper, we use Kinetics-400 as our main video action dataset, due to its popularity.

2.2. Background Cues in Action Recognition

It is widely studied that in the task of object recognition from an image, the background does affect the outcome of the A.I. model [5, 12]. For example [12] explains that machine learning models tend to ignore objects that are in an unusual position, e.g., an elephant in a bedroom. [5] explains this as a 'shortcut learning', where deep learning models tries to perform the task without learning the intended solution.

Similarly, action recognition in video is no different. It is well believed that the background can be a strong cue for the action recognition. E.g., A.I. model will consider a video to be *Playing Tennis*, not from seeing the human body movement, but from seeing the tennis court on the background. There has been an early work [9] that shows that action recognition can be done possible using both background and objects. Some recent works [4, 10] gives and machine learning model that can mitigate the background influence when recognizing a model.

Our work is closest to Mimetics Dataset [16], a video action dataset, where the action is performed in a normal background without using related object, as if the person is 'miming'. The dataset offers 713 videos in 50 classes of Kinetics-400. However, as Mimetics were collected from YouTube videos, they have limited samples of videos with most of them being low-resolution. Moreover, it is arguable if removing an object can still be considered as an action. E.g., if we remove 'cake' from *Eating Cake* video, can the video still be considered as eating a cake? Our work is different from Mimetics as we offer high-quality diverse videos by manually recording the dataset, while keeping the object intact, only recording on different backgrounds.

3. WeirdKinetics-300

In this section we explain our novel dataset, **WeirdKinetics-300**, where we explain our collection methodology and the statistics of the dataset.

3.1. Collection Methodology

Vocabulary Among the set of action classes in the Kinetics dataset, we choose a subset of the classes to focus on. We do not use all 400 classes as it is infeasible to work on all 400 videos for the time being. To make our 10 action classes to be well representative of diverse human actions, we chose 10 action classes according to the following criteria.

- Actions have to be related to their background
- Actions should be from a variety of categories
- Avoid similarity

Thus, we chose four sports action classes and six daily life action classes. Among two of the four sports action

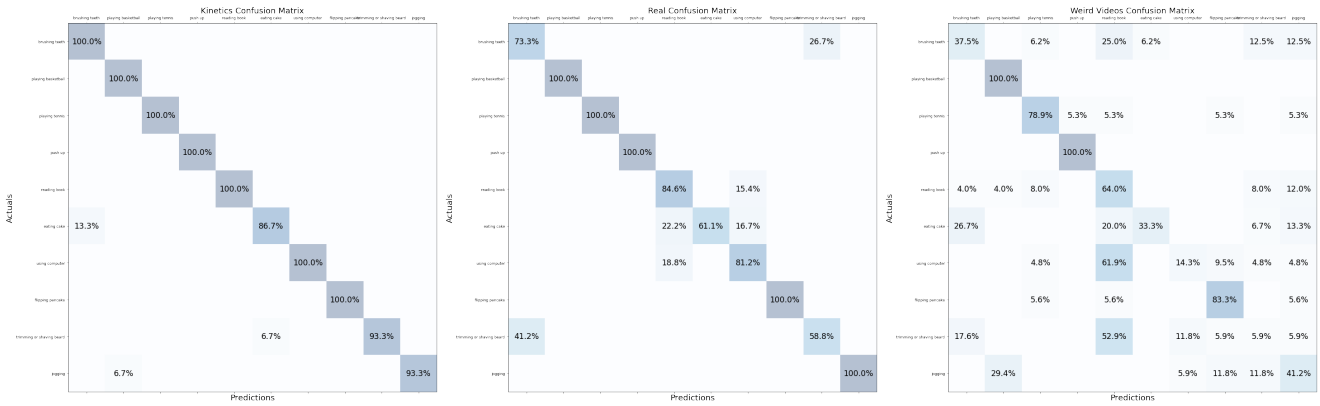

Figure 2. Figure above shows 3 confusion matrices that we have extracted from our experiments. Horizontal axis shows predicted labels while the vertical axis shows the actual labels. Thus, the machine learning model can be seen more accurate if the confusion matrix is close to a diagonal matrix. **Left**: Confusion matrix drawn with original Kinetics-400 validation set. **Middle**: Confusion matrix drawn with our data set involving backgrounds related to the actions. **Right**: Confusion matrix drawn with our data set involving backgrounds not related to the actions.

classes were object and background related actions (*Tennis*, *Basketball*), and the other two were focused on body movement (*Push Ups*, *Jogging*). Daily life action classes included ones related to bathrooms (*Brushing Teeth*, *Shaving Beard*), kitchens (*Eating Cake*, *Flipping Pancake*), and desks (*Using Computer*, *Reading Book*).

Video All of our videos are 10 seconds or longer, and recorded with a phone camera (Samsung Galaxy Note 9 and Samsung Galaxy S10) Half were recorded in the "expected background", and other half in the "weird" background, with at least 15 videos on each set. We try to keep the overall video style as close as with Kinetics-400 dataset.

4. Experiments

For the machine learning model, we have used I3D [3] using ResNet-50 [6] as a backbone. I3D is one of commonly used machine learning models in video action recognition. Instead of training from scratch, we have used off-the-shelf available trained model. Specifically, we have used PyTorchVideo [1], a machine learning package maintained by Meta (Facebook).

The off-the-shelf machine-learning model can only take 8 frames at a time (i.e., a single clip consists of 8 frames). Following the same frame sampling strategy as the available pre-trained model, we have extracted the videos to 30-FPS frames, and fed 8 frames with stride of 8, thus the single clip being roughly 2-seconds long. However, as our videos are 10 seconds long, we feed multiple clips of stride 16, and average the model output. This is a conventional clip sampling strategy.

While the Kinetics-400 pre-trained model can predict within 400 action classes, but our dataset only have 10, thus we need to limit the model to output score within our selected 10 action classes. To do this, we simply ignore the output score for other actions.

We used Python programming language, using PyTorch [11] as our machine learning package. We used Google CoLab as a computing resource for all of our experiments.

4.1. Results and Discussion

In this session, we discuss about our experiment results and debates after them.

4.2. Accuracy

	Kinetics-400	WeirdKinetics-300	
Video Dataset	Validation Set	Real	Weird
Accuracy	97.33%	85.80%	55.80%

Table 1. Table above compares accuracy of the three datasets used in the experiment.

Table 1 tabulates accuracy of the three different datasets. The Kinetics dataset showed the highest accuracy, while the Weird dataset showed the lowest. We assume that Kinetics-400 was the highest because they had the most similar style to the training dataset. In other words, as the model was trained with Kinetics training set, it is largely 'familiar' with Kinetics validation set, thus showing the highest accuracy. In Real, although not as good as Kinetics, since it is recorded with intended background of each action, they

still show relatively high accuracy. In Weird, because of the dataset showing unusual backgrounds for each action, they show the lowest accuracy.

However, accuracy alone cannot fully describe why the model was failing in the Weird dataset. To show that the cause of the failure is due to the background, we have to perform deeper analysis.

4.3. Confusion Matrix

A confusion matrix (or table) tabulates for each actual label (row) we show the percentage of the predicted labels (columns). For example, see the middle matrix of Figure 2, specifically the first row. The row indicates among all the videos of *brushing teeth*, only 73.3% were classified as *brushing teeth* while the rest were classified as *trimming or shaving beard.*

Figure 2 shows three confusion matrices of our experiment. As explained, the left-most matrix is confusion matrix drawn from the Kinetics validation set, and except for a few mistakes, shows good results overall. The middle matrix involves real backgrounds, and they too showed relatively good results and few notable mistakes. The first mistake was that the model confused *trimming or shaving beard* with *brushing teeth*, because the two actions involve the same background (bathroom). The second mistake was made between *reading book*,*eating cake*, and *using computer* as they all happen indoors, and usually sitting on a table or desk. Overall, Real still shows good results, especially on the actions where the body movement is unique (e.g., *push up*) or the unique background (e.g., *playing tennis*). Using Weird dataset shows confusion in many of the actions, except *playing basketball* and *push up*, which has accuracy of 100%. When it comes to *push up*, we suspect the high accuracy is due to the fact that training videos had diverse backgrounds in *push up* videos, while *push up* has very unique action compared to other 9 actions (*push up* is the only action performed lying down). Because *playing basketball* videos all have a person dribbling a basketball, we suspect the basketball was another strong hint for the model, thus even with non-related background (not basketball courts, gyms, etc.), the model was able to perform very well.

4.4. Accuracy per Background

Session 4.2 does mention the significance of backgrounds, but it's mainly about confusion between classes. It doesn't give much information about the relationship between the videos' surroundings and accuracy. In session 4.3 we focus more on backgrounds to find out if certain backgrounds cause inaccuracy and vice-versa.

We looked at the recorded videos and grouped the backgrounds into 5 separate categories. The categories are: outdoors, bathroom, home, indoors, and cafe. Figure 3's X-axis lists the 5 background categories mentioned above, and its

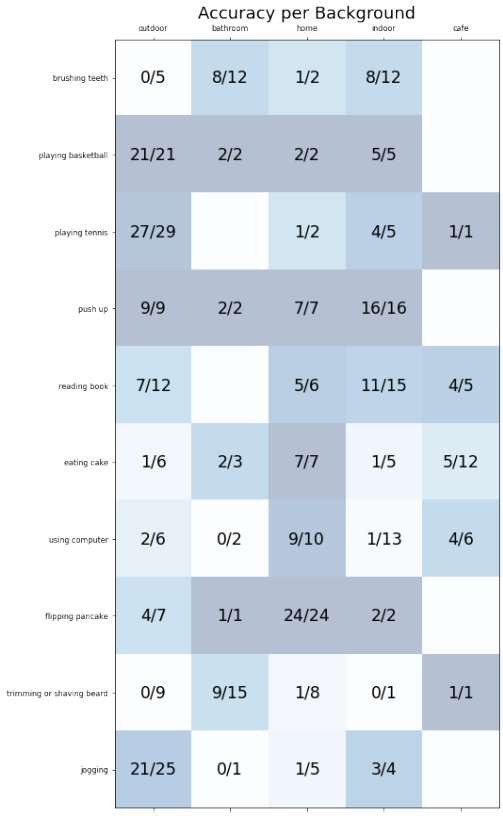

Figure 3. We show accuracy of the trained machine-learning model grouped by the background type. For example, top-left most cell indicates that *brushing teeth* videos recorded outdoors, has shown no accurate predictions among all 5 videos.

Y-axis lists 10 actions we've recorded. Each cell contains the number of correct predictions out of the total videos. Blank cells are where no such videos exist.

On actions that don't rely much on its environment, but focused on the actual movement of the subject (*playing basketball*, *playing tennis*, *push up*, *flipping pancake*), the model showed high accuracy in all backgrounds. However, actions closely related to its backgrounds (*brushing teeth*, *using computer*) had significant accuracy gap in each background. The model shows a near-perfect accuracy on videos in *Using computer* class, with 9/10 at home background, but in different backgrounds such as indoors or bathroom, the model struggles to accurately predict even a single video. We suspect that, when the action recognition model is analyzing a video, the main focus for finding the action *using computer* would be a computer on a desk. When given backgrounds without a desk (outdoors, bathroom, indoors) the overall accuracy was low. Similarly, *brushing teeth* and *trimming or shaving beard* videos show a relatively higher accuracy in bathrooms, but almost zero recognition accu-

racy on videos with an outdoor background.

While we have shown that the actions focused on the person's movement (*playing basketball*, *playing tennis*, *push up*, *flipping pancake*) show high accuracy regardless of surroundings, but other actions, backgrounds do have a significant role in accurate video recognition, as the actions' characteristics heavily rely on its background.

4.5. Qualitative Results

In this section we select some of the sample from the dataset and investigate why the machine learning model has confused, and predicted a different action.

Figure 1 shows examples from the dataset. The left shows samples with unrelated background (thus the model has confused with another action), and the right shows samples with related background.

The left video for *flipping pancake* was recorded in a parking lot. Although both *flipping pancake* videos have the subject flipping a pancake, its background causes a lot of confusion. Because the subject is standing outdoors, and *flipping pancake* is usually done in a kitchen indoors, the machine learning model confused the left video for jogging. Similarly, when the machine learning model was shown the *brushing teeth* video on the left, it mistook it for jogging because its background was outside. The left video for *playing tennis* was recorded in an indoor gym. Although both *playing tennis* videos have the subject swinging a tennis racket, its background causes a lot of confusion. Because the subject is standing indoors, and *flipping pancake* is usually done in a tennis court outdoors, the machine learning model confused the left video for *Push Ups*. The left video for *jogging* was recorded outdoors in a park. Although both *jogging* videos have the subject running, its background causes a lot of confusion. Because the subject is running in a park and holding a basketball, and *jogging* is usually done in a wide, open, space, the machine learning model confused the left video for *basketball*.

5. Limitation and Future Work

There was a limitation to the dataset we have collected. Our dataset is small, with only 300 videos, while the Kinetics have around 100k videos, significantly larger than our dataset. Video backgrounds weren't evenly distributed, especially ones with actions closely related to backgrounds. For example, 21 out of 30 playing basketball videos were recorded outside, and 24 out of 30 flipping pancake videos were recorded indoors at home. The videos also lacked distribution. They all had the same person and were recorded in the same location (Seoul). Also, because of the coronavirus, sport related backgrounds, such as swimming pools and beaches were hard to reach. We hope to collect more diverse and larger version of our dataset that can more accurately show the experiment results, and unveil more issues of machine learning in video action recognition task.

We have also seen limitations to our experiment design. We were aware that the machine learning model focused on backgrounds, but couldn't figure out why. All we did was determine whether the model was confused or not. Because using multiple machines were computationally expensive, we used only one machine learning model. If given enough time and computational machines, we want to experiment with various models and see the overall tendency. For now, we experimented with only one machine learning model, due to the models being computationally expensive. If given enough time and computational machines, we would like to experiment with a variety of models and find out the overall tendency.

6. Conclusion

In this paper We have experimented with a machine learning model and created a "weird background" dataset. The results showed that the machine learning model does rely significantly on backgrounds to determine a video's action, and showed low accuracy on actions heavily related to their backgrounds. With this experiment, we have created the first dataset with a differing background. We hope that A.I. researchers will use our dataset, and help create an A.I. that doesn't only focus on a video's surroundings, but actually sees what the person is doing.

References

[1] Pytorchvideo · a deep learning library for video understanding research. 3

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 3

[4] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[5] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics hu-

man action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[9] Li-Jia Li and Li Fei-Fei. What, where and who? classifying events by scene and object recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 2

[10] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018. 2

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 3

[12] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. 2

[13] Xiaolin Song, Cuiling Lan, Wenjun Zeng, Junliang Xing, Xiaoyan Sun, and Jingyu Yang. Temporal–spatial mapping for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):748–759, 2019. 2

[14] 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. 1

[15] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 1

[16] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *arXiv*. 2