Run, Climb, Bike: Detecting Different Actions With Residual Neural Networks

Most of the decisions we make every day are based on perceived actions. Identifying actions allow us to interpret what is happening in the environment around us. The actions of people in our environment can be a signal to whether the surrounding area is safe or if it is a high-risk area. As well as in everyday life, the importance of recognizing actions emerges with videos as it becomes a common means of communication. Distinguishing the content of the videos from easily accessible sources, such as YouTube, can lead to the usage of it in monitoring security cameras. Security cameras are currently monitored by a human observer, which leads to a potential bottleneck in security monitoring capabilities. Being able to recognize actions through machine learning would increase the number of actions identified and allow for the ability to understand how people behave in specific situations amongst individual people, as the first step of understanding any subject is to identify it.

There have been several recent studies that used the Human Motion DataBase (HMDB 51), a large video database with fifty-one distinct action categories for human motion recognition. Several studies were able to reach above 80% accuracy but did so with different methods. With an 82.1% accuracy, Deep networks with Temporal Pyramid Pooling (DTPP) and pre-training on Kinetics, a large-scale video dataset, were used by learning the video level representation rather than the frame-level feature (Zhu, J., Zhu, Z. & Zou, 2018). With an 81.3 % accuracy, SVM Pooled (SVMP) descriptor was used to by formulating a binary classification problem of separating good feature and bad features with a discriminative classifier. (Wang, Cherian, Porikli, & Gould, 2018). With an 80.9 % accuracy, Pose moTion (PoTion) was used by first running a state-of-the-art human pose estimation using the human joints as key points and then extracts heatmaps for the human joints in each frame. The PoTion representation was obtained by temporally aggregating these probability maps (Choutas, Weinzaepfel, Revaud, & Schmid, 2018). With an 80.7 % accuracy, Two-Stream Inflated 3D ConvNets (I3D) with pre-training on the Kinetics dataset, which builds upon state-of-the-art image classification architectures but inflates their filters and pooling kernels into 3D (Carreira & Zisserman, 2017). This study’s algorithm uses a residual neural network (ResNet), which uses blocks to redirect the input and accumulate the information learned from the previous layer. Doing so will improve the learning and accuracy (He, Zhang, Ren & Sun, 2016). For this study, a ResNet will be used to differentiate between three categories of running, climbing stairs, and riding a bicycle from HMDB 51 that is expected to meet the 80% accuracy rate, similar to the previous studies.

As the data from HMDB 51 is in the form of short video clips, the videos were converted from videos to frames with the dimensions resized to 240 by 320. Each category from the training samples had ten videos with a total of thirty videos and 2271 frames overall, while each category from the testing samples had five videos with a total of fifteen videos and 1126 frames overall. To improve the accuracy, every three frames were grouped together to be consecutively inputted with a learning rate of .0001, a batch size of twenty, and twenty-five number of epochs. These variables were chosen by trial-and-error as single frames had a 40% accuracy with single frames and a learning rate of .001 was overfitting the accuracy for both training and testing samples.

Using our residual network, the network able to correctly classify 99% (Figure 1) of the training samples for three consecutively stacked frames and 80.4% (Figure 4) of the testing samples for three consecutively stacked frames in roughly 30 minutes.

Figure 1 Figure 2

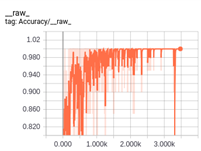
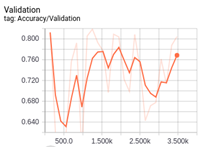
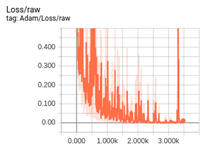


Figure 3 Figure 4



The network was able to classify all the training samples. As seen in Figure 2, and 3, there are instances where the accuracy of the training samples is equal to one. Since obtaining such a high accuracy is unlikely, the training samples most likely overfitting. On the other hand, the network had more difficulty determining the testing samples than the training samples of running, climbing stairs, and riding a bicycle. But even so, the 80.4% of validation accuracy for the testing samples were similar to previous studies. Our study did slightly worse than previous studies, however, this could be due to our study using only three out of fifty-one categories for the training and testing samples. It could also be improved by saving the frames to images after it is converted from video to frames. Overall, our study shows evidence that the method of using a ResNet to differentiate between actions can match up the accuracy of previous studies.

**References**

Carreira, J., & Zisserman, A. (2017, July). Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on* (pp. 4724-4733). IEEE

Choutas, V., Weinzaepfel, P., Revaud, J., & Schmid, C. (2018, June). PoTion: Pose MoTion Representation for Action Recognition. In *CVPR 2018*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Wang, J., Cherian, A., Porikli, F., & Gould, S. (2018, March). Video representation learning using discriminative pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1149-1158).

Zhu, J., Zhu, Z., & Zou, W. (2018, August). End-to-end video-level representation learning for action recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*(pp. 645-650). IEEE.

<https://github.com/mpcrlab/DeepLearningFall2018/blob/RachelWong/Final/RWong_3DiffActions.ipynb>