BRAC
UNIVERSITY
Inspiring Excellence

Project Title: Human Action Recognition With CNN By Following Object Detection and Background Extraction.

Group Members:

Student ID	Student Name	Section
18301253	Murad Hasan	01

Date of Submission: 8 April, 2022

Research paper summary

T. Lima, B. Fernandes and P. Barros, "Human action recognition with 3D convolutional neural network," 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2017, pp. 1-6, doi: 10.1109/LA-CCI.2017.8285700.

According to the authors, human action recognition (HAR) is a significant challenge in computer vision because of constraints such as background clutter, hazy images, view point, lighting, and appearances. As data storage technology advances at a breakneck pace, processing vast amounts of data is no longer a barrier to technological advancement. Additionally, the use of video cameras for various applications such as surveillance and traffic monitoring has expanded significantly, with Human Action Recognition becoming more common for classifying videos via computation. Computer vision technologies are capable of resolving this type of classification challenge by classifying various human actions. Anomaly detection could serve as an excellent illustration of the critical nature of Human Action Recognition (HAR).

The authors of the paper explored a 3D Convolutional Neural Network (CNN) architecture for classifying human activities in a dataset. The major contributions of this study are to experiment with the number of frames required to recognize a human action and to use this architecture to extract the spatial temporal aspects of an activity. Additionally, the authors assessed the suggested method's processing time for various input frame counts and resolutions. This 3D CNN Neural Network is capable of calculating both spatial and temporal data. As the action in human action films changes rapidly, it is critical to calculate the spatial and temporal variables that aid in determining the motion. Additionally, the 3D CNN architecture employs a 3D kernel to perform the convolutional calculation throughout the model's creation. This kernel can convolutionally extract features from several frames, which is a substantial difference between a 2D and 3D model. Additionally, the experimental result for the UCF50 dataset has been displayed at a variety of input resolutions. The experiment findings indicate that it performs well when the input resolution is between 48*48 and 64*64. According to the authors, in order to achieve a greater improvement in accuracy, the input resolution must be raised. Finally, when compared to other proposed models, this proposed model achieves 97.61% accuracy, which is a significant improvement over other models.

In this paper, a confusion matrix for the suggested model is provided, with some values indicated outside the diagonal that are predicted incorrectly. Furthermore, the writers did not provide an in-depth comparison with comparable studies. Biking, Swing, SalsaSpin, PoleVault, WalkingWithDog, and JumpingJack are among the classes that have earned less than 60% accuracy, with the swing and PoleVault having the worst accuracy of all. Because some of the backdrop elements in these classes are the same, there is less precision. Also, according to the authors, it is critical to classify and monitor items in order to improve the accuracy of the results.

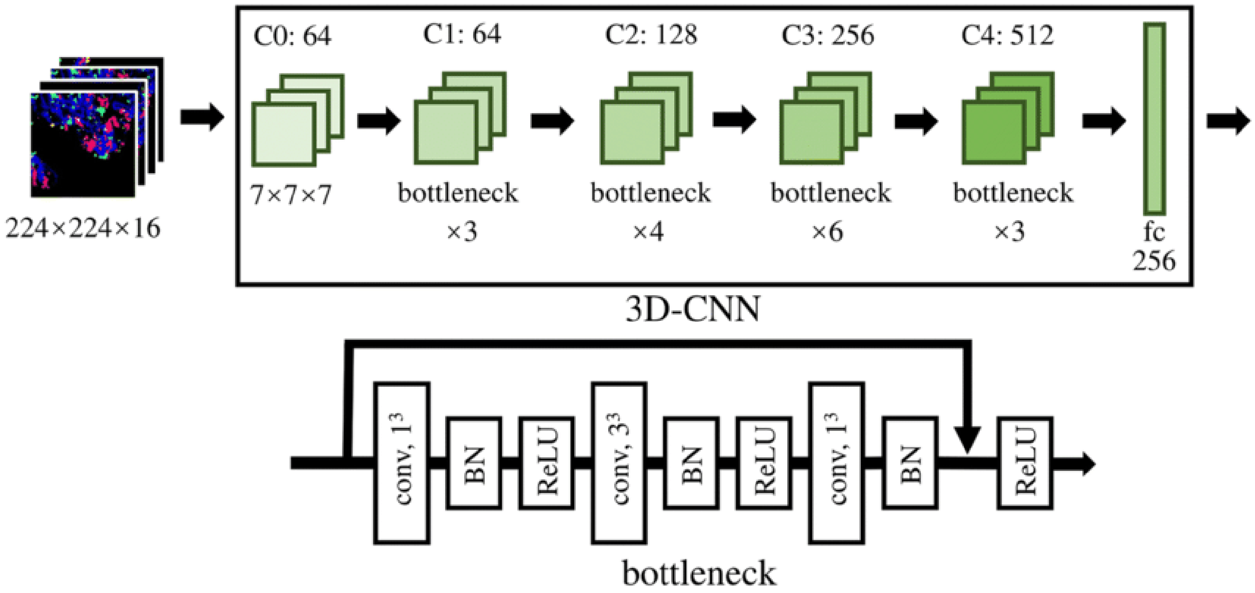
The paper was chosen because human action recognition poses some difficulties, and identifying human activity has become a difficult task as a result of technological improvement. Additionally, we have outlined the model's limits and how they might be solved through the use of specific pre-processing approaches. Recognizing human action in recordings can be an extremely useful technique for identifying any type of crime or anomaly.

Human Action Recognition (HAR) has a big impact on all kinds of activities, natural and unnatural. If there is no computation, it is difficult to identify any type of crime from a surveillance camera. As a result, this assignment can be used to identify any type of crime.
Chron et al. [1] employed different elements of the body to determine the posture of an action. To recognize the motion of the action, the authors used the optical flow technique. They next used a cross-matching technique to classify the activity by comparing the raw photos to the motion they had recognized.
As a result, this 3D CNN model for identifying human action is also applicable to human body position detection. We can come to the conclusion that this strategy can be applied to a variety of other tasks.

Project description

As described by the author, The reason for the lower accuracy is that some of the background elements in these classes are the same, hence our goal is to eliminate the background elements using pre-processing techniques. The accuracy of the new output may improve if we remove the background pieces and train it again in a 3D CNN architecture. We propose using Mask RCNN to recognize our target items in order to eliminate the background elements. The reason for adopting this target is that our reasoning suggests that background factors are the primary cause of poor accuracy in some classes. Furthermore, if we discover the primary items before training the model, the accuracy increase by specific objects can be bigger. The secondary purpose is to identify human pose estimation and create a hybrid model that incorporates additional features collected from the pose estimation. The data gathered from pose estimation will assist the model in achieving high accuracy.

We will also be applying a 3D CNN model, called ResNet, to identify the human activity in addition to the traditional CNN model. In recent years, ResNet has increased in prominence as a classification method for images. Over a million parameters are used in this model.



3D ResNet Model

The fundamental concept of ResNet is the introduction of a so-called "identity shortcut connection" which bypasses one or more levels.

For the purpose of training our model, we will be using the UCF5o dataset. This dataset has a total size of 3GB. This is one of the most widely used datasets for human action recognition, and it is available for free download. It has 50 classes of human action where each of the classes contains more than 100 videos on average. The frames will be extracted from our dataset, and any background elements will be removed before we begin processing the data. Furthermore, we will maintain the 240*240 resolution of the images.

The method we explained earlier ResNet will be implemented for our project. The method will be implemented by us. However, we can use some online resources to demonstrate the accuracy of the model and to compare with the previous model.

We will demonstrate a matrix for many classes to evaluate the outcome of our approach. In the classes where the old model failed, we hope to gain better accuracy. We would try to get more than 80% accuracy, as the old model achieved 60% accuracy in those classes. We'd also show the result by putting it to the test on some random videos to see if our method can appropriately classify the action.

References:

[1] G. Cheron, I. Laptev, and C. Schmid, “P-cnn: Pose-based cnn features ´ for action recognition,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3218–3226