3D Convolutional Neural Networks for Human Action Recognition

Feb 28, 2023

Project Team: Opara Anita (V00996592), Sekhon Anuinder (V01022326)

Description of Paper

IEEE Citation for Paper: S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

Mission: The paper extends the use of Convolutional Neural Networks to 3D inputs by proposing a 3D CNN model for human action detection by considering the features extracted from the real environment video dataset. It compares the performance to the existing 2D CNN systems.

Problem Addressed

The chosen paper addresses the problem of human action recognition in surveillance videos using real-world data from airport surveillance as context. It proposes a deep learning model based on neural networks that is independent of the assumptions about the surrounding environment in which the video was taken and provides enhanced performance over the existing models that address similar situations. This algorithm can be applied in engineering tasks that require the analysis of objects in motion such as in security for video surveillance, to study shopping behavioural analysis for personalized recommendations, detecting panic situations in crowd, and in the medical industry for assistive care [1].

Here,3D convolution is used to extract spatial and temporal features from the video input dataset which is used for the CNN model. Multiple features were extracted from the same input by adjusting the kernel size to produce different feature maps. The 3D CNN model used in this paper takes the input of seven video frames of size 60 by 40. The initial kernel(filter) was of the size of 7*7*3 (7*7 for the spatial dimension and 3 for the temporal dimension). The paper aims to identify three human actions (CellToEar ,ObjectPut and Pointing) and classifies all other actions as negative. C++ was used for implementing the model.

We believe that the research requires legal permissions of the geographical area to be implemented in the real world as it requires accessing the surveillance videos. Along with this, it requires taking into consideration the cultural beliefs of that region as well as the privacy of individuals because humans that appear in the video dataset have to be informed about their data being collected.

Solution and Dataset

To address the problem of human recognition in surveillance videos, this paper has proposed the use of a deep learning model called 3D Convolutional Neural Networks(3D-CNN). This algorithm has been primarily selected for two reasons – invariancy to the surrounding environmental conditions and the ability to consider motion information along both spatial and temporal dimensions using video analysis.

Previously, 2D Convolutional Neural Networks have been applied to similar problems where each video image is treated as a still frame to extract the features and then study classifiers on the extracted features.

This 3D-CNN model takes into consideration multiple features as extracted by applying convolutional operators without making assumptions about the viewpoint and scale of video and claims to provide the enhanced performance along all three participating categories. This way the paper claims to capture more useful data from the dataset.

This model was trained and evaluated on 2 datasets: the TRECVID 2008 development dataset and the KTH dataset using supervised learning. For the TRECVID dataset, the paper did not state the number of samples used for training although it highlighted that a large number of labelled samples were used. In the KTH dataset,16 of the 25 inputs were randomly selected for training while the remaining 9 were used for testing. The TRECVID dataset was used for testing as part of an event detection system while the KTH dataset was used for testing the model independently.

TRECVID 2008 development data dataset contains a 49 hour long videos taken real-time from 5 different cameras on five days (20071101, 20071106, 20071107, 20071108, and 20071112),at the London Gatwick Airport. The Airport is a place that has a high amount of human traffic as each data frame contained multiple different humans carrying out various random actions.

References

[1] Gupta N, Gupta SK, Pathak RK, Jain V, Rashidi P, Suri JS. Human activity recognition in artificial intelligence framework: a narrative review. Artif Intell Rev. 2022;55(6):4755-4808. doi: 10.1007/s10462-021-10116-x. Epub 2022 Jan 18. PMID: 35068651; PMCID: PMC8763438.