Chapter 1
Video Classification Using 3D Convolutional Neural Network

K. Jairam Naik
https://orcid.org/0000-0002-6332-418X
National Institute of Technology, Raipur, India

Annukriti Soni
National Institute of Technology, Raipur, India

ABSTRACT

Since video includes both temporal and spatial features, it has become a fascinating classification problem. Each frame within a video holds important information called spatial information, as does the context of that frame relative to the frames before it in time called temporal information. Several methods have been invented for video classification, but each one is suffering from its own drawback. One of such method is called convolutional neural networks (CNN) model. It is a category of deep learning neural network model that can turn directly on the underdone inputs. However, such models are recently limited to handling two-dimensional inputs only. This chapter implements a three-dimensional convolutional neural networks (CNN) model for video classification to analyse the classification accuracy gained using the 3D CNN model. The 3D convolutional networks are preferred for video classification since they inherently apply convolutions in the 3D space.

INTRODUCTION

Convolutional neural networks are considered and verified as a prime algorithm for image and video classification. Presently, they deliver an effective result for

DOI: 10.4018/978-1-7998-2795-5.ch001

image identification and dissection. Because of the upright outcomes in images, they are considered for video acknowledgment. In the field of computer vision, the semantic incident apperception encounters the fascinating one at present. It denotes that a succession of humanoid gesticulations in a video is signified as motions. For example, boxing, driving, etc. Especially, human action apperception is a consequential research area due to sundry applications such as video scrutiny and client comportment examination. Action apperception focuses on detecting certain activities from a video frame and to relegate those frames consequently. General and robust models can be provided by the Convolution neural networks for video or image apperception predicaments with insignificant manual work and can be elongated to numerous types of situations easily.

In chapter, the emphasis is on 3D inputs utilizing the convolutional neural network. The utilization of 3D convolution sanctions to capture timely-based three-dimensional data from videos by captivating successive frames into consideration is the key task. The importance is to analyze the 3D CNN performance for video classification, compare it with manual techniques and to genuinely comprehend it's working.

The primary motivation for the development of effective video retrieval systems is an explosion in the volume of media data over wireless or internet networks. The secondary motivation is the increasing admiration of imaging devices such as digital camera and increasing proliferation of image data over communications networks. The emergence of new consumerism where media technologies meet consumer needs. A video retrieval system is typically an application where users can construct queries such as "Show medical history for brain injury cases with CT scans similar to this one". The problem at the core of such applications is finding images that are visually similar. There are many ways to compute visual similarity but not all are efficient. Simple and classic measures such as Euclidian distance (or) Manhattan distance only compute the difference between pixel values while totally ignoring visual queue. The current algorithm for video retrieval is not accurate and computationally expensive and hence to deploy in real life becomes tedious. Hence, a robust model is designed, implemented and deployed that meets all the real-time constraints and is more efficient and accurate when compared to the traditional existing models.

LITERATURE REVIEW

For the classification of large scale videos using the Convolutional Neural Network, about 1 million sports videos were considered by [A. Karpathy et al, June 2014] from YouTube. That data was treated with the 2D convolutional neural network. The focus was to consider the large sports dataset for classification purposes, although