Applied Soft Computing 145 (2023) 110575

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Fast 3D-graph convolutional networks for skeleton-based action recognition

Guohao Zhang, Shuhuan Wen *, Jiaqi Li, Haijun Che

Department of Key Lab of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao, 066004, China

ARTICLE INFO

Article history:
Received 20 July 2022
Received in revised form 14 June 2023
Accepted 20 June 2023
Available online 29 June 2023

Keywords:
Action recognition
Human skeleton
Graph convolutional
Knowledge distillation

ABSTRACT

Research on human action recognition based on skeletons has received much attention. But most of the research focuses on improving the model's generalization ability, while ignoring significant efficiency issues. This leads to developing heavy models with poor scalability and cost-effectiveness in practical use. This paper, we investigate the under-studied but practically critical recognition model efficiency problem. To this end, we present a new Fast Recognition Distillation (FRD) model learning strategy. Specifically, FRD trains a lightweight recognition neural network structure that can be quickly executed at a low computational cost. It can be achieved by effectively disseminating the identification probability information of the teacher network to the lightweight network. We call the probability information of the teacher network as soft-target, and FRD can learn more potential information from soft-target. In addition, we also used a particular loss function for soft-target. Through the FRD network, while basically maintaining the recognition accuracy, we minimized the network structure. Extensive experiments on the two large-scale datasets, NTU-RGBD and Kinetics-Skeleton, demonstrate that our model (FRD) is more lightweight and refined than others. Therefore, our model FRD is efficient.

1. Introduction

Human action recognition has always been a significant research issue in the computer vision domain and an essential part of human recognition for machines. It is a process marking the video streams of human behavior as the corresponding behavior category with multiple high-level applications like competent healthcare, intelligent education, intelligent video surveillance [1], multimedia information retrieval [2,3], human–computer interaction [4,5], interactive entertainment [6] and so on. With the rapid development of computer hardware, we can easily collect multi-modal human motion data, including RGB, depth, and human skeleton data. In contrast, video data like RGB can be easily affected by those external factors unrelated to behaviors, for example, the shooting environment, illumination, the pedestrians clothing and its texture, etc. Therefore, skeleton-based action recognition has become a new research direction. Initially, all body joint positions in each frame are encoded as feature vectors for pattern learning [7–9]. However, with the development of graph neural networks, the skeleton action recognition based on graph neural network (*GNN*) [10–12] has been an important research topic recently.

GNN, a newly emerging technique, differs from the traditional neural network in that the processed data is based on the graph, a non-Euclidean data. The non-Euclidean data is used in many scientific research fields, such as social networks in social sciences, functional networks in brain imaging and protein structure networks in the biochemistry field. Similarly, human skeleton diagrams are also irregular data. A method was proposed to construct a skeleton graph with vertices as joints and edges as bones to capture the correlation between human joint points. It uses the graph convolutional network (*GCN*) to extract relevant features and further develops the spatio-temporal *GCN (ST-GCN)* [13] to learn spatio-temporal characteristics. When the *ST-GCN* model the video sequence, it gives all frames the same weight, despite that most actions are finished within 50 frames. Consequently, Tang et al. [14] found that many *CNN* models based on skeleton often consider each frame unequally important. Previous models failed to focus on more representative frames, so Reinforcement Learning (*RL*) was proposed to choose frames and divide the model into frame distillation network (*FDNet*) and graph convolutional neural network (*GCNN*). Then useful frames obtained from the *FDNet* are input into the GCNN for action recognition. Although the effectiveness has been improved, the computation of the two network models is vast. In addition, we found that *ST-GCN* can only extract the parts that are naturally connected to the human body joints. However, for some long-distance interactive points, the extraction failed. For example, when people are walking, their hands and feet are used together,

* Corresponding author.
E-mail addresses: zgh@stumail.ysu.edu.cn (G. Zhang), swen@ysu.edu.cn (S. Wen), leejq@stumail.ysu.edu.cn (J. Li), hjche@ysu.edu.cn (H. Che).

https://doi.org/10.1016/j.asoc.2023.110575
1568-4946/