Computer Vision and Image Understanding xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Exploiting deep residual networks for human action recognition from skeletal data

Huy-Hieu Pham[*,a,b], Louahdi Khoudour[a], Alain Crouzil[b], Pablo Zegers[c], Sergio A. Velastin[d,e]

[a] Centre d'Études et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement (CEREMA), Toulouse 31400, France
[b] Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse, UPS, Cedex 9, Toulouse 31062, France
[c] Aparnix, La Gioconda 4355, 10B, Las Condes, Santiago, Chile
[d] Department of Computer Science, Applied Artificial Intelligence Research Group, University Carlos III de Madrid, Madrid 28270, Spain
[e] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK

ARTICLE INFO

Keywords:
3D Action recognition
Deep residual networks
Skeletal data

MSC:
68T10
68T45
68U10

ABSTRACT

The computer vision community is currently focusing on solving action recognition problems in real videos, which contain thousands of samples with many challenges. In this process, Deep Convolutional Neural Networks (D-CNNs) have played a significant role in advancing the state-of-the-art in various vision-based action recognition systems. Recently, the introduction of residual connections in conjunction with a more traditional CNN model in a single architecture called Residual Network (ResNet) has shown impressive performance and great potential for image recognition tasks. In this paper, we investigate and apply deep ResNets for human action recognition using skeletal data provided by depth sensors. Firstly, the 3D coordinates of the human body joints carried in skeleton sequences are transformed into image-based representations and stored as RGB images. These color images are able to capture the spatial-temporal evolutions of 3D motions from skeleton sequences and can be efficiently learned by D-CNNs. We then propose a novel deep learning architecture based on ResNets to learn features from obtained color-based representations and classify them into action classes. The proposed method is evaluated on three challenging benchmark datasets including MSR Action 3D, KARD, and NTU-RGB+D datasets. Experimental results demonstrate that our method achieves state-of-the-art performance for all these benchmarks whilst requiring less computation resource. In particular, the proposed method surpasses previous approaches by a significant margin of 3.4% on MSR Action 3D dataset, 0.67% on KARD dataset, and 2.5% on NTU-RGB+D dataset.

1. Introduction

Human Action Recognition (HAR) is one of the key fields in computer vision and plays an important role in many intelligent systems involving video surveillance, human-machine interaction, self-driving cars, robot vision and so on. The main goal of this field is to determine, and then recognize what humans do in unknown videos. Although significant progress has been made in the last years, accurate action recognition in videos is still a challenging task due to many obstacles such as viewpoint, occlusion or lighting conditions (Poppe, 2010).

Traditional studies on HAR mainly focus on the use of hand-crafted local features such as Cuboids (Dollár et al., 2005) or HOG/HOF (Laptev et al., 2008) that are provided by 2D cameras. These approaches typically recognize human actions based on the appearance and movements of human body parts in videos. Another approach is to use Genetic Programming (GP) for generating spatio-temporal descriptors of motions (Liu et al., 2012). However, one of the major limitations of the 2D data is the absence of 3D structure from the scene. Therefore, single modality action recognition on RGB sequences is not enough to overcome the challenges in HAR, especially in realistic videos. Recently, the rapid development of depth-sensing time-of-flight camera technology has helped in dealing with problems, which are considered complex for traditional cameras. Depth cameras, *e.g.*, Microsoft Kinect ™sensor (Cruz et al., 2012; Han et al., 2013) or ASUS Xtion (ASUS, 2018), are able to provide detailed information about the 3D structure of the human motion. Thus, many approaches have been proposed for recognizing actions based on RGB sequences, depth (Baek et al., 2017), or combining these two data types (RGB-D) (Wang et al., 2014) , which are provided by depth sensors. Moreover, they are also able to provide real-time skeleton estimation algorithms (Shotton et al., 2013) that help to describe actions in a more precise and effective way. The skeleton-based representations have the advantage

* Corresponding author at: Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse, UPS, Cedex 9, Toulouse 31062, France.
E-mail address: huy-hieu.pham@cerema.fr (H.-H. Pham).

https://doi.org/10.1016/j.cviu.2018.03.003
Received 29 September 2017; Received in revised form 15 January 2018; Accepted 6 March 2018

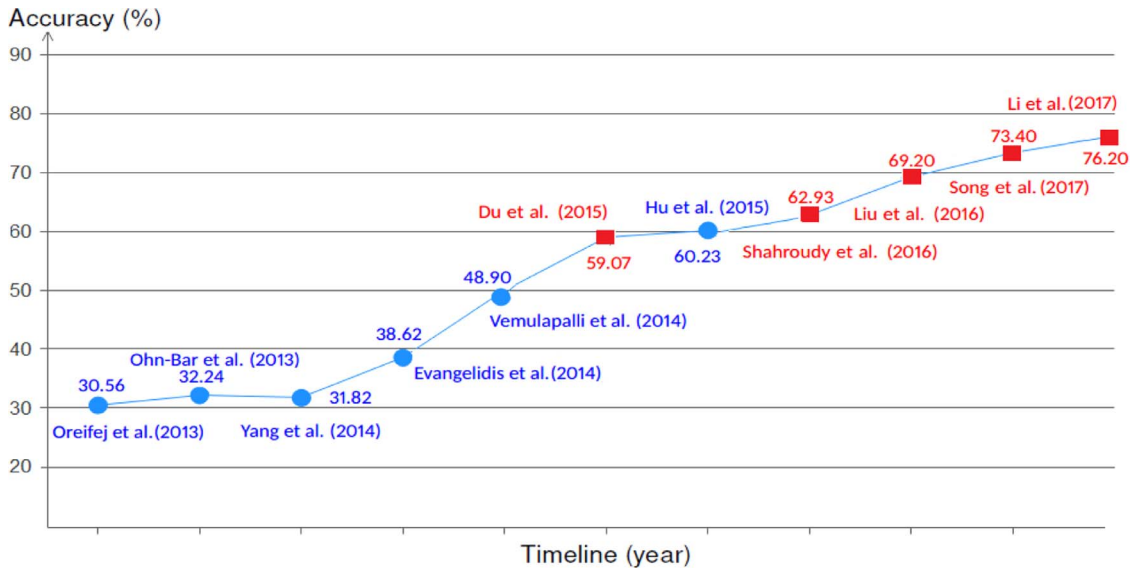

Fig. 1. The recognition performance of hand-crafted and deep learning approaches reported on the Cross-View evaluation criteria of NTU-RGB+D dataset (Shahroudy et al., 2016). The traditional approaches are marked with circles (Evangelidis et al., 2014; Hu et al., 2015; Ohn-Bar and Trivedi, 2013; Oreifej and Liu, 2013; Yang and Tian, 2014). The deep learning based approaches are marked with squares (Du et al., 2015; Li et al., 2017; Liu et al., 2016; Shahroudy et al., 2016; Song et al., 2017).

of lower dimensionality than RGB/RGB-D-based representations. This benefit makes action recognition systems become simpler and faster. Therefore, exploiting the 3D skeletal data provided by depth sensors for HAR is a promising research direction. In fact, many skeleton-based action recognition approaches have been proposed (Chaudhry et al., 2013; Ding et al., 2016; Vemulapalli et al., 2014; Wang et al., 2012; Xia et al., 2012b).

In recent years, approaches based on Convolutional Neural Networks (CNNs) have achieved outstanding results in many image recognition tasks (Karpathy et al., 2014; Krizhevsky et al., 2012). After the success of AlexNet (Krizhevsky et al., 2012) in the ImageNet competition (Russakovsky et al., 2015), a new direction of research has been opened for finding higher performing CNN architectures. As a result, there are many signs that seem to indicate that the learning performance of CNNs can be significantly improved by increasing their depth (Simonyan and Zisserman, 2014b; Szegedy et al., 2015; Telgarsky, 2016). In the literature of HAR, many studies have indicated that CNNs have the ability to learn complex motion features better than hand-crafted approaches (see Fig. 1). However, most authors have just focused on the use of relatively small and simple CNNs such as AlexNet (Krizhevsky et al., 2012) and have not yet fully exploited the potential of recent state-of-the-art very deep CNN architectures. In addition, most existing CNN-based approaches use RGB, depth or RGB-D sequences as the input to learning models. Although RGB-D images are informative for action recognition, however, the computation complexity of these models will increases rapidly when the dimension of the input features is large. This makes models become more complex, slower and less practical for solving large-scale problems as well as real-time applications.

In this paper, we aim to take full advantages of 3D skeleton-based representations and the ability of learning highly hierarchical image features of Deep Convolutional Neural Networks (D-CNNs) to build an end-to-end learning framework for HAR from skeletal data. To this end, all the 3D coordinates of the skeletal joints in the body provided by Kinect sensors are represented as 3D arrays and then stored as RGB images by using a simple skeleton-to-image encoding method. The main goal of this processing step is to ensure that the color images effectively represents the spatio-temporal structure of the human action carried in skeleton sequences and they are compatible by the deep learning networks as D-CNNs. To learn image features and recognize their labels, we propose to use Residual Networks (ResNets) (He et al., 2016) – a very deep and recent state-of-the-art CNN for image recognition. In the hope of achieving higher levels of performance, we propose a novel deep architecture based on the original ResNets, which is easier to optimize and able to prevent overfitting better. We evaluate the proposed method on three benchmark skeleton datasets (MSR Action 3D Li et al., 2010; Kinect Activity Recognition Dataset - KARD Gaglio et al., 2015; NTU-RGB+D Shahroudy et al., 2016) and obtain state-of-the-art recognition accuracies on all these datasets. Furthermore, we also point out the effectiveness of our learning framework in terms of computational complexity, the ability to prevent overfitting and to reduce the effect of degradation phenomenon in training very deep networks.

The contributions of our work lie in the following aspects:

- Firstly, we propose an end-to-end learning framework based on ResNets to effectively learn the spatial-temporal evolutions carried in RGB images which encoded from skeleton sequences for 3D human action recognition. To the best of our knowledge, this is the first time ResNet-based models are applied successfully on skeletal data to recognize human actions.
- Secondly, we present a novel ResNet building unit to construct very deep ResNets. Our experiments on action recognition tasks prove that the proposed architecture is able to learn features better than the original ResNet model (He et al., 2016). This architecture is general and could be applied for various image recognition problems, not only the human action recognition.
- Finally, we show the effectiveness of our learning framework on action recognition tasks by achieving the state-of-the-art performance on three benchmark datasets including the most challenging skeleton benchmark currently available, whilst requiring less computation.

The rest of the paper is organized as follows: Section 2 discusses related works. In Section 3, we present the details of our proposed method. Datasets and experiments are described in Section 4. Experimental results are shown in Section 5. In Section 6, we discuss classification accuracy, overfitting issues, degradation phenomenon and computational efficiency of the proposed deep learning networks. This section will also discuss about different factors that affect the recognition rate. Finally, Section 7 concludes the paper and discusses our future work.