Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-12091-z

3DFCNN: real-time action recognition using 3D deep neural networks with raw depth information

Adrián Sánchez-Caballero[1] · Sergio de López-Diz[1] · David Fuentes-Jimenez[1] · Cristina Losada-Gutiérrez[1] · Marta Marrón-Romera[1] · David Casillas-Pérez[2] · Mohammad Ibrahim Sarker[1]

Received: 25 March 2021 / Revised: 10 June 2021 / Accepted: 3 January 2022

Abstract
This work describes an end-to-end approach for real-time human action recognition from raw depth image-sequences. The proposal is based on a 3D fully convolutional neural network, named 3DFCNN, which automatically encodes spatio-temporal patterns from raw depth sequences. The described 3D-CNN allows actions classification from the spatial and temporal encoded information of depth sequences. The use of depth data ensures that action recognition is carried out protecting people's privacy, since their identities can not be recognized from these data. The proposed 3DFCNN has been optimized to reach a good performance in terms of accuracy while working in real-time. Then, it has been evaluated and compared with other state-of-the-art systems in three widely used public datasets with different characteristics, demonstrating that 3DFCNN outperforms all the non-DNN-based state-of-the-art methods with a maximum accuracy of 83.6% and obtains results that are comparable to the DNN-based approaches, while maintaining a much lower computational cost of 1.09 seconds, what significantly increases its applicability in real-world environments.

Keywords 3D-CNN · Action Recognition · Depth Maps · Real-time · Video-surveillance

1 Introduction

In the computer vision field, human action recognition (HAR) has gained a great importance in recent years, mainly due to its multiple applications in the study of human behavior, safety or video surveillance, which has attracted the attention of many researchers [5, 31, 52, 75, 86].

This work has been partially supported by the Spanish Ministry of Science and Innovation under projects EYEFUL (PID2020-113118RB-C31/C33), by the Community of Madrid under project CONCORDIA (CM/JIN/2021-015) and by the University of Alcal under project ARGOS+ (PIUAH21/IA-016)

✉ Cristina Losada-Gutiérrez
cristina.losada@uah.es

Extended author information available on the last page of the article.

There are several works in the literature whose aim is recognizing human actions. The first proposals were based on the analysis of RGB sequences [55, 69, 72, 96], using different datasets [4]. The release of RGB-D cameras [16, 58], that in addition to a color image provide a depth map (in which each pixel represents the distance from the corresponding point to the camera), has allowed the appearance of numerous works that address HAR using RGB-D information [1, 2, 12, 27, 39], 3D skeleton data [27, 35, 87, 91], or both [70] with acceptable results. In addition, several new datasets including RGB-D information for action recognition have been made available to the scientific community [97].

Most of the previously cited proposals provide good results in controlled conditions, however, they present problems in scenarios with a high degree of occlusions. Besides, the use of color information (RGB or RGB-D) implies the existence of data that allows the users' identification, so problems related to privacy may appear. The use of depth cameras [34], that obtain just information about the distance from each point of the scene to the camera by indirectly measuring the time of flight of a modulated infrared signal, allows to preserve people privacy, since it is not possible to know their identity from those data. Another advantage to consider is that depth cameras do not require additional lighting sources, as they include an infrared lighting source. Thus, depth maps are also widely used in different works for action recognition [79–81, 88].

In recent years, the improvements in technology and the appearance of large-scale datasets [42, 53, 59] have led to an increase in the number of works that use deep learning for different computer vision applications, such as semantic segmentation [22, 48, 93], or HAR from RGB sequences [13, 24, 56, 76], RGB-D data [9, 21, 84], depth maps [38, 46, 78, 89] or 3D skeletons [26, 27, 29, 36]. Most of these works are based on using Recurrent Neural Networks (RNNs), or Long Short Term Memory (LSTM). Besides, more recent works combine these networks with Generative Adversarial Networks (GAN) [66], that have demonstrated their efficacy in different computer vision tasks, such as face alignment [49] or face editing [50].

All these works provide good accuracy but with a high computational cost, which mostly prevents its operation in real-time applications.

Despite the numerous works dealing with the recognition of actions, HAR still remains an open issue in real scenarios, with open problems such as the different viewpoints in videos, changing lighting conditions, occlusions, etc.

The present paper describes 3DFCNN, an approach for real-time action recognition which performs the feature extraction and action classification steps in a single stage, see Fig. 1.

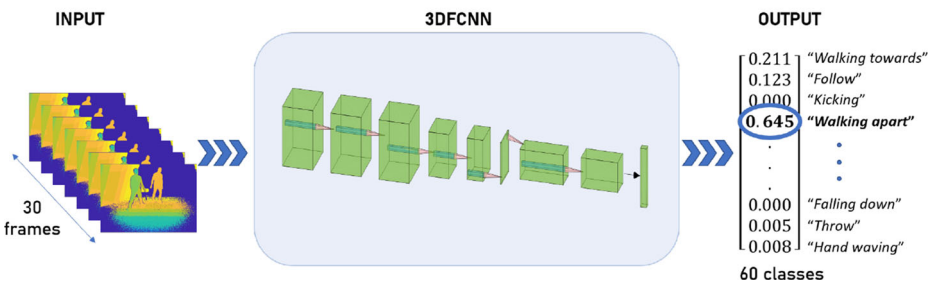


Fig. 1 General scheme of the proposed 3DFCNN for real-time action recognition

3DFCNN is based on a 3D fully convolutional neural network (3D-CNNs), which only uses the raw depth information provided by a depth or RGB-D camera to recognize the actions. It is an end-to-end trainable model, composed by a first phase to extract the main spatial and temporal features, using 3D convolutions and pooling layers, and a final softmax layer for obtaining the detected action. As Section 5 details, 3DFCNN outperforms state-of-the-art methods with a much shorter computational cost due to its reduced architecture optimised to reach the real-time operation. Furthermore, the use of depth data that not allows recognizing people preserves their privacy.

Both training and testing stages have been carried out with the widely used "NTU RGB+D Action Recognition Dataset" [42, 59], made available to the scientific community by the ROSE Lab of the Nanyang Technological University of Singapore. This dataset has been chosen because it provides a large number of videos, both with RGB and depth information, and it allows to compare, through the experimental results, different works of the state-of-the-art. Furthermore, to test the robustness of the 3DFCNN and compare it to that of other previous works, it has also been evaluated using two multiview 3D action datasets: NorthWest-UCLA Multiview Action 3D [73] and UWA3D Multiview Activity II [53].

We want to remark the main contributions of the present work in the following points:

1. Our 3DFCNN approach has been optimized to reach a very good performance in terms of accuracy working in real-time. As we will see in the Experimental Section, 3DFCNN obtains accuracy rates close to the best state-of-the-art methods whose implementations require lots of computations which slow-down the classification. Our 3DFCNN surpasses all of them in terms of speed being the first real-time approach for action detection with these accuracy rates.
2. A careful design of the 3DFCNN architecture has been done for fulfilling the speed requirements. Fully connected layers have been replaced by convolutional layers which has considerably reduced the computation time and number of parameters allowing real time speed rates, while maintaining good accuracy rates.
3. Our 3DFCNN is an end-to-end approach, which means that it does not need any preprocessing or initialization step to detect actions. It receives a depth video as input and provides the classified action. Many of the action detection methods that use only depth information work with a preprocessed data such as those based on precomputed 3D-Skeleton [26, 43, 61] or Depth Dynamic Images (DDI) [78, 89].
4. The present work makes a thorough review and comparison with the best state-of-the-art method. 3DFCNN is closed in accuracy to [78, 89] which are the most comparable methods of the state-of-the-art which work also with raw depth data. It obtains around a 7% under the rates obtained by both methods. However, it is much faster, 1.09s of computation time, compared to the previous one which need around 1 minute of computation time to obtain their results. We also show a comparison with 3D-Skeleton-based action detection methods, which appeared before.

The rest of this paper is organized as follows: in Section 2 the main related works are presented and analyzed, next, in Section 3 the architecture of the neural network proposed is described. Then, in Section 4, the training method used is explained. Subsequently, Section 5, includes the main experimental results obtained, and finally, Section 6 includes the main conclusions of the work, as well as some possible lines of future work.

2 Related works

As stated in the introduction, multiple proposals for HAR have been developed during the last decade, based on different visual technologies. In this section, we analyze the most interesting in order to be compared with the proposal hereby described. The main HAR works can be classified into three groups depending on the technology used: RGB, RGB-D or depth images [65, 95]. Below, it is presented a brief analysis of the RGB and RGB-D-based works, and a more in-depth study of those one which only use depth data for HAR. Figure 2 summarizes the state-of-the-art in HAR using Deep Neural Networks (DNN), with which the proposed method is compared.

2.1 Color based methods (RGB and RGB-D)

The first works in HAR were based on the use of RGB sequences [3, 8, 52, 55]. These works require a feature extraction system followed by a classification process for action recognition. More recently, the improvements in technology and the availability of large-scale datasets have led to an increase in the number of related works that use RGB data and are based on deep learning techniques, including both, supervised [13, 62, 76, 77], and more recently, unsupervised approaches [56].

When combining the RGB channels of an image with its depth information, RGB-D images are obtained. In recent times, analysis of these images or videos are sought due to the availability of real-time inexpensive depth sensors that provide rich 3D structural data. Thus, this has motivated the proposal of numerous works based on combining RGB and depth data for HAR [12, 14, 19, 20, 39, 40, 45, 100].

Also, the promising results achieved by deep learning methods in computer vision applications have encouraged their utilization on RGB-D images and videos for HAR [9, 21, 28, 32, 41, 74, 84], mostly based on the use of CNNs [1] and Recurrent Neural Networks (RNNs).

2.2 Depth-based methods

Besides, regarding to HAR proposals that rely on just depth information, also several modalities of depth data have been used in the related literature. Some studies use raw depth maps [46, 78–81, 88, 89], whereas others extract specific 3D information from them, as a preprocessing task, like joint positions of human body (skeleton data) [26, 37, 43, 61, 68, 92, 98]. In both cases, deep learning methods have replaced conventional methods to process this data, especially when a large scale dataset is available.

2.2.1 HAR using skeletons

The main contributions to action recognition using 3D skeletons have been focused on input data representations and improvements of deep learning methods. Although there have been some research proposals for action recognition based on CNN using 3D skeleton data with good results [26, 29, 36], most recent studies use RNNs and, in particular, variations of the long short-term memory (LSTM) unit, which solves the gradient vanishing and exploding problems [15, 25]. In addition, LSTM networks are able to learn long-term dependencies, which is essential in action recognition problems. Song et al. proposed in [68] a modified LSTM with a joint selection gate for creating an spatio-temporal attention model. In a

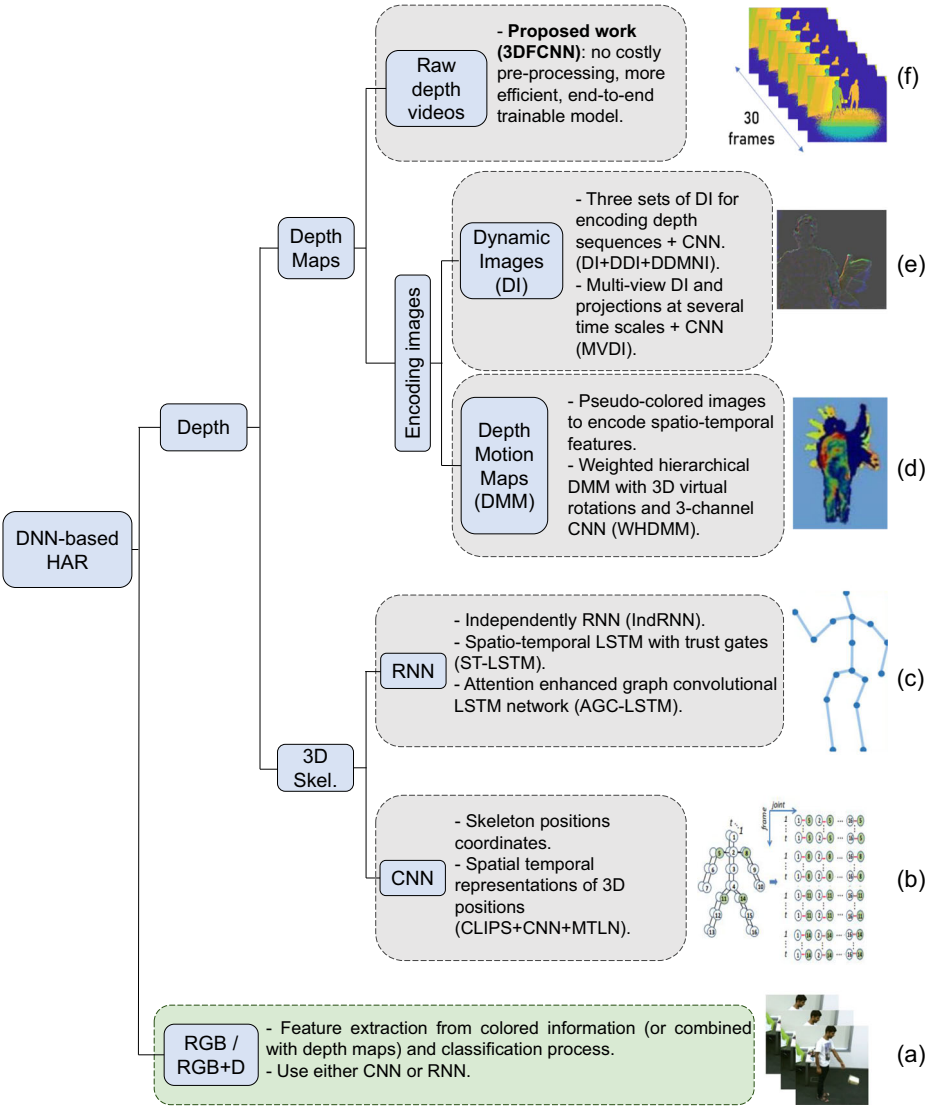


Fig. 2 Summary of the state-of-the-art in DNN-based HAR. *a)* (image from [59]) refers to the RGB-based methods detailed in Section 2.1. The rest of the methods are depth-based, see Section 2.2. *b)* ([26]) and *c)* ([61]) are 3D-skeleton-based methods explained in Section 2.2.1. The first category uses CNNs and the second one RNNs. *d)* ([81]), *e)* ([78]) and *f)* ([59]) are depth map-based methods, deeply explained in Section 2.2.2

similar way, in [43] it was introduced a trust gate for the LSTM and a 3D skeleton data representation in a tree-like graph was proposed, whereas other proposals built a view adaptive LSTM scheme [98] to solve the problem of sensitive viewpoint variations in human action videos.

Li et al. in [37] presented a new type of RNN: the independently RNN (IndRNN), that improved previous results in 3D skeleton-based action recognition. In IndRNN, neurons

inside a layer are independent among each other but connected across layers, favouring the stacked-layers scheme.

Recent approaches use an extension of the graph convolutional networks (GCN) in order to allow the network learning spatio-temporal features more efficiently. Yan et al. in [92] proposed a spatio-temporal graph convolutional network (ST-GCN) which automatically learns both the spatial and temporal patterns from 3D skeleton data by, firstly, transforming it into a graph.

Similarly, Si et al., in [61], achieved state of the art results by using an attention enhanced graph convolutional LSTM network (AGC-LSTM). This model presents also a temporal hierarchical architecture and takes into account the co-ocurrence relationship between spatial and temporal domains.

2.2.2 HAR using raw depth maps

Besides, other researchers used raw depth maps, avoiding some known problems related to 3D skeleton position extraction as spatial information loss in images, extraction failures and sensitivity to pose variations. Moreover, by using raw depth maps, the entire 3D structure information of a scene can be used for recognition. The similarity of depth images with RGB ones allows to transfer all the knowledge from RGB-based action recognition proposals to the depth modality. Thus, the success of CNN methods in RGB-based recognition methods makes reasonable to transfer these studied techniques to the depth domain, and that is what most recent studies have done. For instance, most works [78–81] have used a pseudocoloring technique by which a depth sequence is encoded into several RGB images (depth motion maps and dynamic images), transforming spatio-temporal patterns to colored properties like textures, as it is explained next.

In addition, some approaches as the ones described in [79–81] apply 3D rotations to point clouds from depth images, to use 3 orthogonal projections, leveraging the 3D structure of depth data and, concurrently, augmenting samples of the training dataset. In these approaches, depth motion maps are generated from these projected sequences using different temporal pooling methods, referred to as rank pooling. Alternatively, Luo et al. in [46] proposed an encoder-decoder framework using an unsupervised LSTM network through 3D flows for motion description.

Similar to depth motion maps, three sets of dynamic images (DI) were proposed in [78, 83] for encoding depth sequences: dynamic depth image (DDI), dynamic depth normal image (DDNI) and dynamic depth motion normal image (DDMNI). These images are constructed using hierarchical and bidirectional rank pooling to capture spatio-temporal features. Therefore, for each video sample, it is used an ensemble of 6 images as input of 6 independent CNNs through a VGG-16 architecture [63].

Also, spatially structured dynamic depth images have been proposed for selecting attributes to be used for action recognition [85]. In this approach, rank pooling has been employed to extract three pairs of structured dynamic images, one at each body part, and with joint level granularity. This aids in retaining the spatio-temporal details as well as in improving the structural particulars at different granularity at various temporal scales. Further, spatially and temporally structured dynamic depth images have been highlighted through hierarchical and bidirectional rank pooling methods in order to derive spatial, temporal and structural characteristics from the depth channel of the image [18].

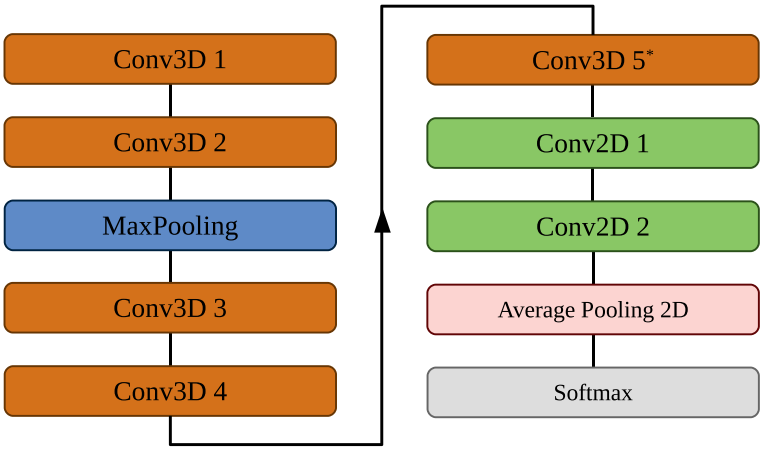
In a similar way, Wu et al. in [88] proposed using depth projected difference images as a dynamic image with a hierarchical rank pooling with 3 pre-trained CNNs, one for each orthogonal projection.

Finally, a multi-view imaging framework (MVDI) was proposed in [89] for action recognition with depth sequences. First, it is employed a human detection network to reduce the action region. A total of 11 different viewpoints are used then for generating projected images at several temporal scales, making thus data augmentation. Also, they propose a novel CNN model where dynamic images with different viewpoints share convolutional layers but have different fully connected layers for each view. The final action classification is made with support vector machines (SVM) with principal component analysis (PCA). All this process results in a method that achieves state of the art on action recognition accuracy on several datasets but with a high computational cost.

3 Network architecture

We propose an end-to-end trainable Deep Neural Network-based (DNN) model for solving the HAR problem with depth maps, named as 3DFCNN. Our proposal is represented by a 3D fully convolutional action recognition architecture, that is very efficient for action recognition tasks. This type of networks, in comparison with the classical convolutional feature extractor followed by fully connected layers, are more efficient, specially in terms of parameters, and avoiding overffiting, which is a well-known problem with fully connected layers. In addition to this, we adapted the network specially for the action recognition task by being very fast in processing the 3D tensors, which are the big flow of information that represents the videos and significantly slows down the network processing. This type of architecture has proven to be very efficient for action recognition tasks in RGB videos [11]. Next, it is described in detail the complete architecture of the used neural network, whose general structure is shown in Fig. 3.

The 3DFCNN network is fed by a sequence of consecutive depth images with size 64×64 pixels, which constitutes its single input. Each sequence has a fixed number of frames, specifically, 30 frames of a video fragment, corresponding to 1 second of a video record (30 fps) showing the execution of a single action. This number of frames has been chosen



*Removes temporal dimension of tensor

Fig. 3 Simplified architecture summary of the proposed 3DFCNN

to achieve a balance between accuracy and computation time. The use of sequences with a reduced number of frames allows decreasing the computational cost and enables real-time performance. Besides, in some works such as in [57] 7-10 frames are enough for recognizing simple actions. However, in the case of NTU RGB+D dataset [59] that includes complex and very long actions (the maximum video length is 300 frames), this sequence length may not be enough to fully cover them. In contrast, the use of longer sequences increases the computational cost, reducing the processing speed. Moreover, in this situation, overlapping may occur between different actions in continuous action videos.

In this paper, there has been made an experimental adjustment of the number of frames, with the aim of achieving a balance between the computational cost and the accuracy of the algorithm. The optimal length has been determined to be 1 second (30 frames) for the input sequences of our deep learning model. It is enough for including significant information for most of the detected actions, without increasing the computation cost. Besides, it reduces the possibility of several actions overlapping in the same video-segment in real sequences, which can include more than one action throughout the video. Since the NTU RGB+D dataset includes videos with very different length, there has been proposed a method for selecting 30 frames for each complete video that is detailed later, in Section 4.

In contrast to 2D CNNs, in which the operations are carried out only on the spatial dimension of input images, in a 3D-CNN, features are extracted by applying 3D convolutional filters on the spatial and temporal dimensions of input videos. This is necessary for the neural network to take into account the context and temporal changes of actions for a better recognition.

Input sequences are processed by first applying two 3D convolutional layers with 32 filters each, padding with zeroes in order to conserve tensor dimensions, and then a dimensionality reduction through a *Max Pooling* layer. Second, another convolutional block of two layers with 64 filters each one, without any padding. Next, the temporal dimensionality is removed with an additional Conv3D layer with 128 non-squared filters. After that, a Conv2D layer with 128 filters precedes a final Conv2D layer with a number of filters that matches classes number in the dataset (60 in case of NTU RGB+D dataset).

Finally, an activation layer *softmax*, noted as S, is used to compress the output vector to real values between 0 and 1, in order to obtain a normalized likelihood distribution, see (1):

$$\begin{aligned} &S\colon \mathbb{R}^{60} \to [0,1]^{60} \\ &a = (a_1, \cdots, a_{60}) \mapsto S(a) = (S_1, \cdots, S_{60}) \end{aligned} \tag{1}$$

The (2) shows the formula applied to obtain the probabilities for each action.

$$S_j = \frac{e^{a_j}}{\sum_{k=1}^{60} e^{a_k}} \quad 1 \le j \le 60 \tag{2}$$

Besides, at the output of each convolutional layer (except for *Conv3D 5* and *Conv2D 2*), a *Batch Normalization* [23] layer and a *LeakyReLU* (*Leaky Rectified Linear Unit* [47]) activation function are included. Batch Normalization helps training the neural network reducing the internal covariate shift. The Leaky ReLU activation function follows the expression $f(x) = x$ if $x \ge 0$ and $f(x) = \alpha x$ if $x < 0$, with $\alpha = 0.3$. This type of function has been chosen instead of the conventional ReLU due to its proven greater efficiency [90], and because it provides the necessary non-linearity to solve the action recognition problem avoiding gradient vanishing problems [33].

In addition, it is also applied a slight dropout regularization technique for helping the model to generalize more and reduce over-fitting problems during training.

Table 1 Network architecture parameters and output tensor sizes for each layer

Layer	Output shape	Parameters
Input	64 × 64 × 30 × 1	–
Conv3D 1	64 × 64 × 30 × 32	kernel=(3, 3, 3) / strides=(1, 1, 1)
Batch Normalization	–	
Activation	LeakyReLU	
Conv3D 2	64 × 64 × 30 × 32	kernel=(3, 3, 3) / strides=(1, 1, 1)
Batch Normalization	–	
Activation	LeakyReLU	
MaxPooling	22 × 22 × 10 × 32	size=(3, 3, 3)
Dropout	0.25	
Conv3D 3	20 × 20 × 8 × 64	kernel=(3, 3, 3) / strides=(1, 1, 1)
Batch Normalization	–	
Conv3D 4	18 × 18 × 6 × 64	kernel=(3, 3, 3) / strides=(1, 1, 1)
Batch Normalization	–	
Activation	LeakyReLU	
Dropout	0.25	
Conv3D 5	18 × 18 × 1 × 128	kernel=(1, 1, 6) / strides=(1, 1, 1)
Reshape	18 × 18 × 128	–
Conv2D 1	8 × 8 × 128	kernel=(3, 3) / strides=(2, 2)
Batch Normalization	–	
Activation	LeakyReLU	
Conv2D 2	8 × 8 × 60	kernel=(1, 1) / strides=(1, 1)
Average Pooling 2D	60	–
Activation	softmax	

A more detailed description of the different layers that form the proposed neural network is shown in Table 1, as well as their output sizes and fundamental parameters.

The efficient network architecture and the input data generation system make possible to reduce the overall computational complexity of the model, allowing a quite fast performance unlike most other action recognition works in the literature. Thus, the number of parameters (400.476) of our proposal is much lower compared to some well-known classification baselines such as VGG16 [64] with a total of 138 million parameters and Resnet50 [17] with a total of 234.355.586 parameters, giving us a briefly idea of how efficient our network is in terms of parameters. It opens the door to the capability of real-time performance [6, 10, 44, 94] and its important applications in video-based health care service, video surveillance or human-computing interaction.

4 Network training

The large-scale NTU RGB+D dataset [59] has been used for training and testing the 3DFCNN. This dataset contains 56 880 video sequences for 60 different human actions. These sequences were recorded using *Microsoft Kinect II* sensors [99] obtaining RGB images, depth maps, 3D skeleton positions and infrared data. In this paper, only depth map

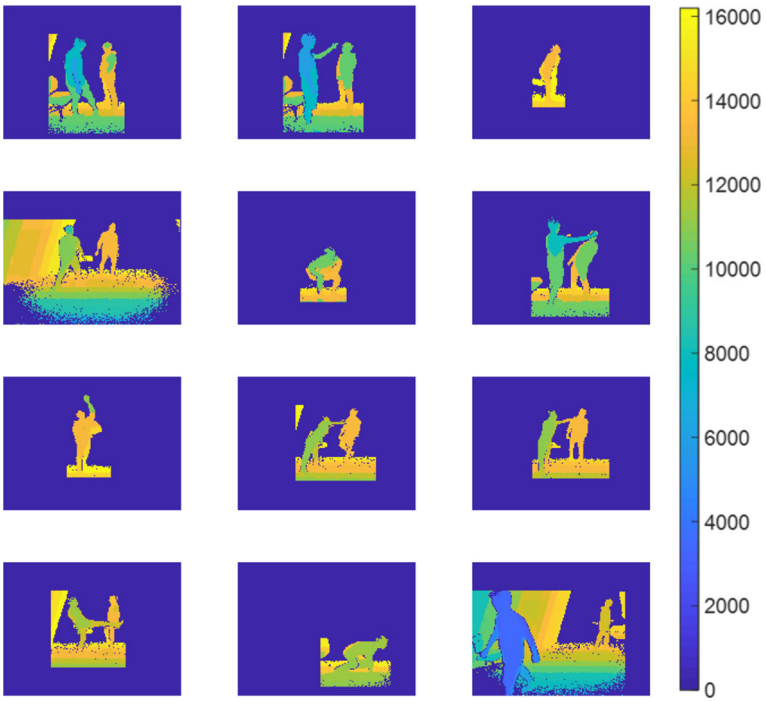


Fig. 4 Examples of masked depth images from different video sequences in NTU RGB+D dataset. As shown, it includes several subjects performing different actions from varying viewpoints

videos are used, which are composed of frames with an original resolution of 512×424 pixels. However, authors of this dataset suggest to use the masked depth maps, which are the foreground masked version of the full depth mask and, therefore, have a much better compression ratio, which facilitates the download and file managing. Some examples of masked frames are shown in Fig. 4.

Furthermore, in order to be able to focus on the HAR issue (without previously detecting people) and reduce the number of pixels without useful information, all images have been cropped to the region of interest in which the action is happening, decreasing the weight of files at the same time. Figure 5 is an illustrative example of this image cropping. This process is performed by a simple code that removes the outermost rows and columns with all elements equal to zero. Finally, the cropped image is re-scaled to fit the 64×64 input size. A bilinear interpolation is applied for adapting the original frame size of the dataset to the input size of our 3DFCNN network.

As it has been explained before, the input of the neural network is composed of 30 frames sequences, so the full input size is $30 \times 64 \times 64$ pixels. Experiments have confirmed the importance of how these 30 frames are selected. Consequently, there is set a data arrangement strategy for training, so as to take into account the dataset properties used. NTU RGB+D dataset contains videos with lengths between 26 to 300 frames. The data generator is specifically modified to optimally select the 30 frames for each sample as follows. When videos are shorter than 30 frames, last frames are repeated backwards until completion. When greater than 30 but shorter than 60 frames, the starting frame is randomly selected inside suitable limits. Finally, when videos are greater or equal to 60 frames the generator

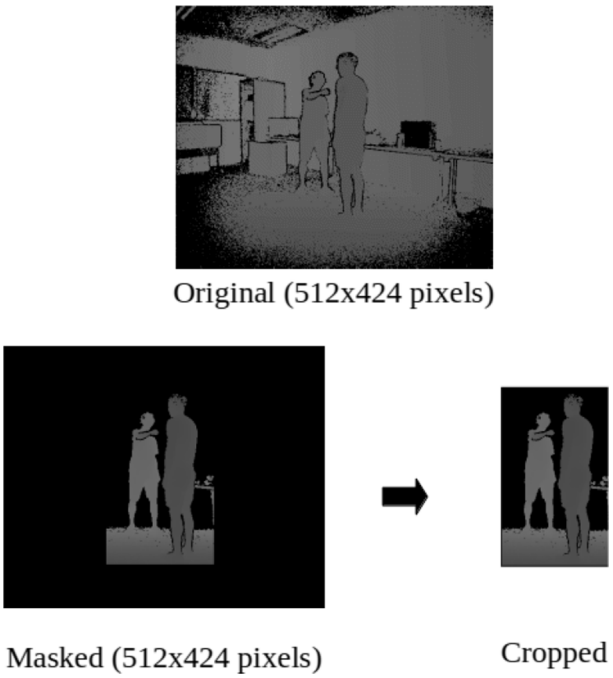
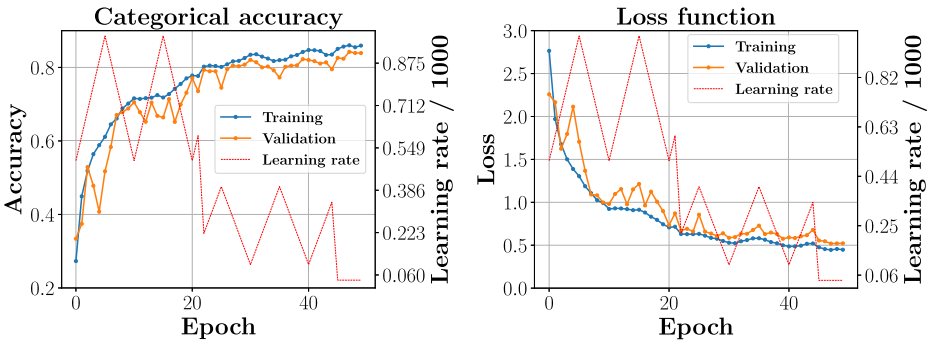


Fig. 5 Example of a full depth map (up) and the process of cropping a masked depth map into a smaller image (down)

randomly selects a suitable starting frame with one-frame-skipping, i.e. taking only odd or even frames. This has two advantages: first, processing the double video length with only 30 frames in case of long actions and second, making data augmentation thanks to random starting point. This strategy has proved to be adequate to successfully extract relevant information from both short and longer videos for a good performance of the classification.

Batch size has been experimentally set to 12, being a reasonable value that makes a good balance between hardware-memory and model generalization. Once the batch size is set, a learning rate range test is performed in order to find which are the best values for this hyperparameter. This test consists in modifying the learning rate value along a wide range and analyzing the loss function behaviour.

A cyclical learning rate schedule is chosen due to its benefits [67], like a faster convergence and a reduced over-fitting. The training procedure consists of a fixed number of epochs in which the network is fed by the batch-structured random training dataset. Once the whole set of batches are used to train the network, an epoch is finished. Then, a new epoch starts again by randomly permuting the training dataset initiating the process. The number of times that this process is repeated is called the number of epochs, which has been empirically optimized as we explains below. Taking into account the range test results, the learning rate is made to oscillate first between 5×10^{-4} and 9.8×10^{-4}, then between 1×10^{-4} and 4×10^{-4} and finally fixed to 4×10^{-5} in the last 5 epochs. The specific curves of learning rate along the training can be seen in Fig. 6a and b, where accuracy and loss functions for training and validation are shown, respectively. For the optimization process, the algorithm *Adam* [30] has been used due to its proven adaptive properties and



(a) Categorical accuracy (b) Loss function

Fig. 6 Training (in blue) and validation (in orange) categorical accuracy and loss function curves. The learning rate schedule is superposed in red

computational efficiency. The training of the neural network has been carried out on a PC with a NVIDIA GeForce GTX 1080 with 8 GB.

Figure 6a and b show the process of training the neural network along 50 epochs. This number of epochs have demonstrated to be enough for the model to reach convergence. It can be seen that, thanks to the applied deep learning techniques (dropout, fully-convolutional and cyclical learning rate), although a large dataset as NTU RGB+D is used, validation and training curves are significantly close to each other, i.e. there does no appears over-fitting. Furthermore, it can be observed that the validation loss slightly diverges (or analogously, the validation accuracy falls) when the learning rate takes the higher values of the cycle. This is a desired effect because it prevents the solution to stay stuck at saddle points and helps reaching better minima.

5 Experimental results and discussion

5.1 Experimental setup

The proposed 3DFCNN has been trained and tested using NTU dataset [59]. This dataset is used in most of the deep learning based approaches for action recognition due to the large number of available videos, which allows training deep neural networks. Furthermore, the use of this dataset facilitates comparison to other state-of-the-art approaches, mainly based on DNNs. To evaluate the robustness of the 3DFCNN, it has also been tested in other two multiview publicy available datasets, with different actions and camera characteristics, as well as a lower number of sequences for training and testing. These three datasets have been selected to ease comparison with other methods, since there are ones of the most used in

Table 2 Characteristics of the three datasets used for testing the proposed 3DFCNN

Dataset	Samples	Classes	Subjects	Sensor	Modalities
NWUCLA	1494	10	10	Kinect v1	RGB+D, 3DJoints
UWA3DII	1070	30			
NTU RGB+D	56880	60	40	Kinect v2	RGB+D, 3DJoints, IR

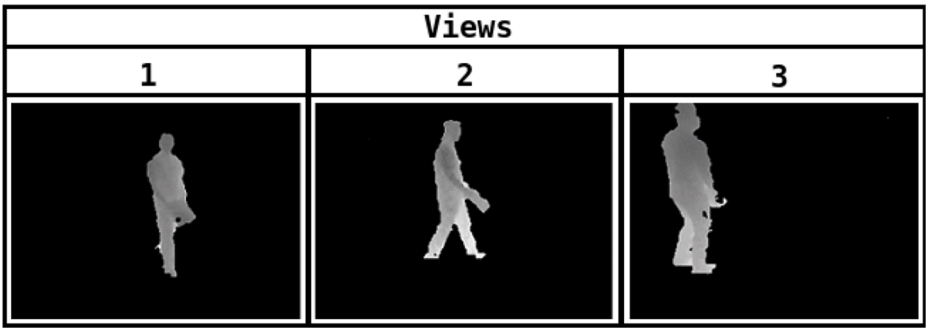


Fig. 7 Northwestern-UCLA dataset sample images

the HAR related scientific literature. The main characteristics of each dataset are briefly described below. Besides, Table 2 shows a summary of these characteristics for allowing comparison.

Northwesternt-UCLA Multiview Action3D dataset (NWUCLA) contains multiview RGB, Depth and 3D joints data acquired using three Kinect v1 cameras in a variety of viewpoints, see Fig. 7.

It includes 10 different actions performed by 10 subjects, with a total of 1494 sequences (518 for view 1, 509 for view 2 and 467 for view 3) with different lengths. In the cross-view setting, the authors propose using two views for training and one view for testing. Figure 7 shows three views of a sample image belonging to this dataset.

UWA3D Multiview Activity II dataset (UWA3DII) [53] was collected using a Kinect v1 sensor, and it includes RGB, Depth and 3D joints, see Fig. 8.

This dataset is focused on cross-view action recognition, and includes 10 subjects performing 30 different human activities, recorded from 4 different viewpoints (frontal, left, right and top view). Each subject performs the same action four times in a continuous manner, and each time the camera is moved to record the action from a different viewpoint. As

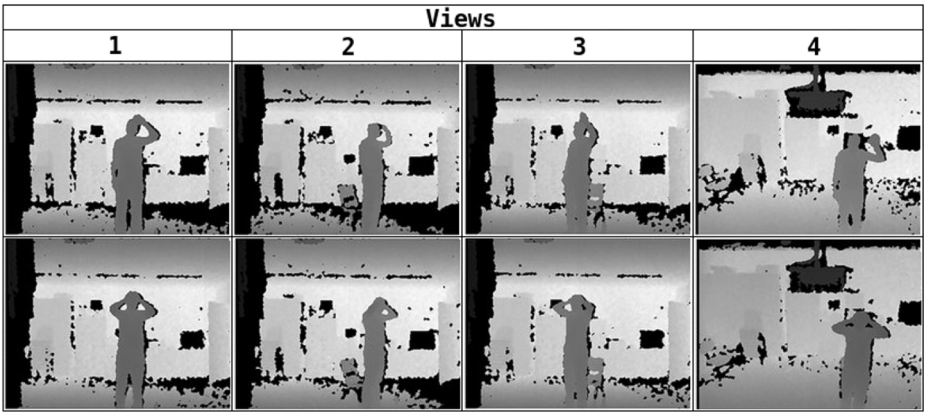


Fig. 8 UWA3DII dataset sample images

a result, there are 1070 sequences. It is a challenging dataset because of varying viewpoints, self-occlusion and similarity between different actions (such as *walking* and *irregular walking* or *drinking* and *phone answering*). The authors propose cross-view action recognition, using the samples from two views for training, and the two remaining views for testing. The complete evaluation includes six different combinations of the 4 views for training and testing, as it can be seen in Table 6. Specifically, the first row of Table 6 indicates the views used for training during the experiments. The second row shows the results testing over the other views. Vertical lines separate the different experiments. Some sample images of this dataset are shown in Fig. 8.

The large-scale **NTU RGB+D dataset** contains 56880 video samples for 60 different actions performed by several subjects. The dataset has been acquired using three Kinect V2 cameras concurrently. For each sample, there are RGB videos, depth maps, IR (Infrared) data and 3D joints.The actions in NTU RGB+D dataset can be organized in three major categories: daily actions, mutual actions, and medical conditions. It is worth highlighting that *mutual actions* (such as *pushing other person*, *hugging other person*, *giving something*, etc.) involve more than one people. Some sample images from this dataset can be seen in Fig. 4. The authors of NTU RGB+D dataset propose two different evaluations [59] for separating data between training and testing:

1. *Cross-Subject* (CS) evaluation in which there are 40 320 training samples with 20 subjects and 16 560 with other 20 different subjects for testing.
2. *Cross-View* (CV) evaluation, with 37 920 sequences from 2 different viewpoints for training and 18 960 from a third camera for testing.

As it can be seen in Table 2, the first two datasets (NWUCLA and UWA3DII) include a reduced number of samples, which makes it difficult training DNNs without overfitting. These two datasets also have a lower number of classes than the NTU RGB-D dataset. Besides, the used sensor and the viewpoints are different. Furthermore, it is worth highlighting that the Kinect v1 used in UWA3DII dataset provides images with a higher amount of noise and measurement errors than Kinect v2, as it can be seen in Fig. 8

Due to the small number of available training samples for UWA3DII and NWUCLA datasets, instead of training the network from scratch for each of them, we have started with the model trained using NTU RGB+D dataset. Then, the three last layers in the network have been fine-tuned using a cyclical learning rate with the same schedule that for the NTU dataset. Besides, since the images provided in the NTU RGB+D dataset are masked, there has been necessary to remove the background for the images belonging to these two datasets.

For the proposal evaluation, there have been used the evaluation protocols proposed for the authors of each dataset. Furthermore, the obtained results are compared to other works evaluated in these datasets. It is worth highlighting that most of the works that are based on DNNs use the large scale NTU RGB+D dataset because of the number of available samples, whereas the works based on non-DNN-based methods are usually evaluated in NWUCLA or UWA3DII.

5.2 Experiments on NTU RGB+D dataset

For the experimental test of the proposal, there has been used the two proposed evaluations in [59] for separating data between training and test for the NTU RBG+D dataset. Before going further in the comparison, Fig. 9 shows the performance of our 3DFCNN method in several action recognition examples of the NTU dataset.

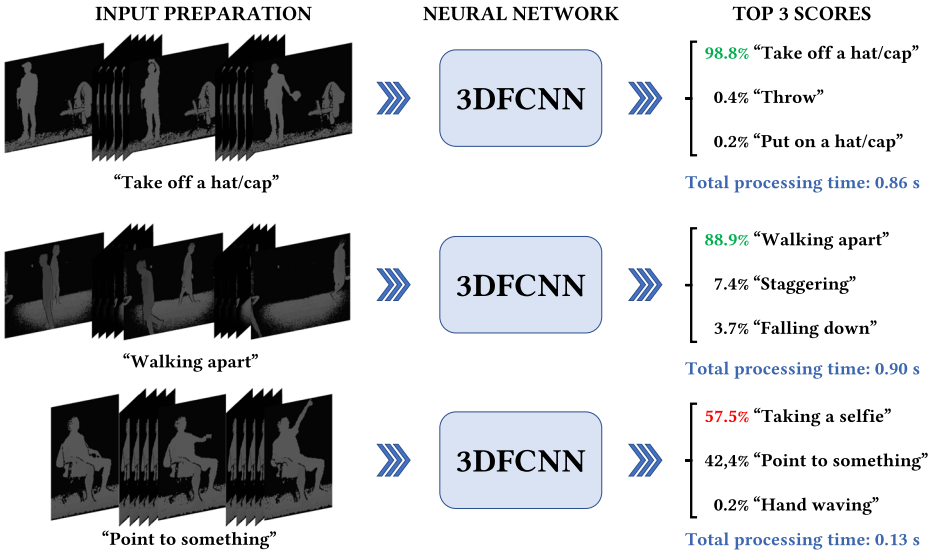


Fig. 9 Three examples of the action recognition by 3DFCNN

Specifically, the actions are "take of a hat/cap", "walking apart" and "point to something". 3DFCNN achieves to clasify the first two actions with high probability 98.8% and 88.9% respectively. However, the action "point to something" was misunderstood by our method which has been relegated to a second position with a probability of 42.4%. Our method chooses the action "taking a selfie" instead with a 57.5% of probability. This action is clearly closed to the true action "point to something". Both are hardly indistinguishable in the depth domain, even for humans. Notice the extremely small amount of time that our method takes to elaborate the predictions in each case, below 1 second which makes it useful for real-time applications.

The method proposed in this paper has achieved an accuracy on NTU RGB+D dataset of 78.13% (CS) and 80.31% (CV) maintaining a very low computational cost, as detailed below, where an analysis of the different per-classes recognition accuracies, as well as a comparison with previous methods in terms of accuracy and computational cost are presented.

The confusion matrix obtained for the NTU RGB+D dataset is shown in Fig. 10. It gives a general view of the model performance (for the CV evaluation), showing that it provides a good classification performance for the 60 different classes.

Table 3 shows the 10 best recognized actions and the 10 most confused action pairs.

As it can be seen, the proposed model can not totally recognize very similar actions like *reading* (mostly confused with *writing*), *writing* (with *type on a keyboard*) or *play with phone/tablet* (with *type on a keyboard*). It is noteworthy that, in general, the proposed model performs really well with actions where small objects are not involved. Nevertheless, where there are small objects which are discriminatory for the action recognition, e.g. phone/tablet, shoes, meals, toothbrush, etc., the model tends to missclassify towards a similar action. The reason for this may be the small input image size that is fed to the neural network (64×64), along with the absence of color and texture of objects.

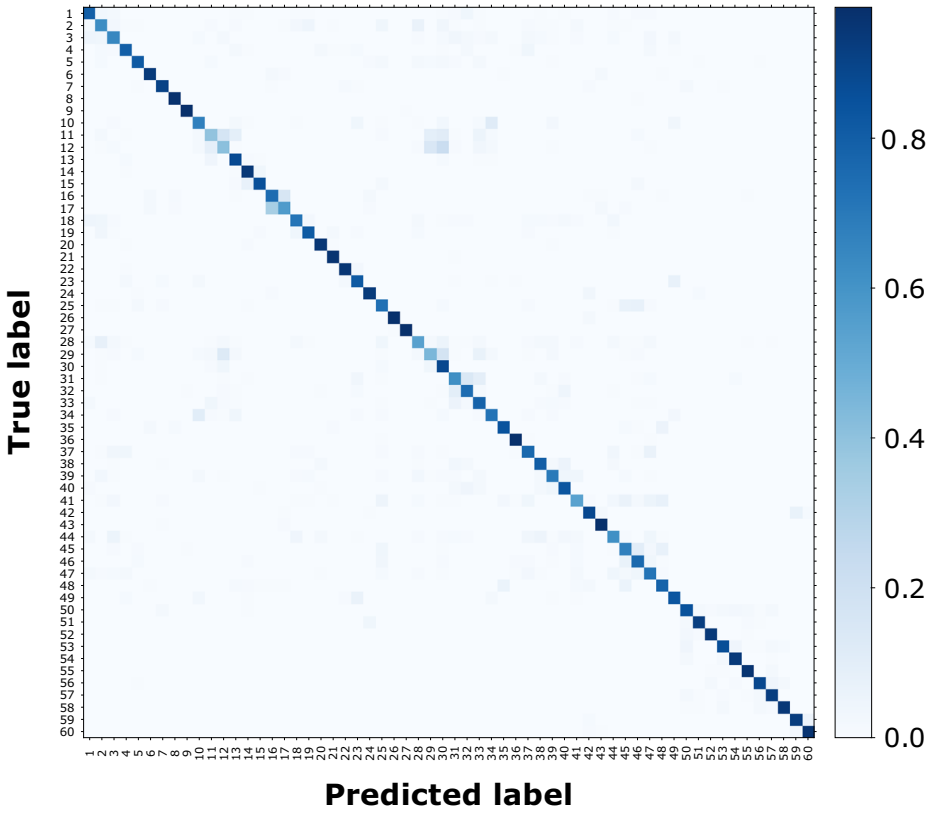
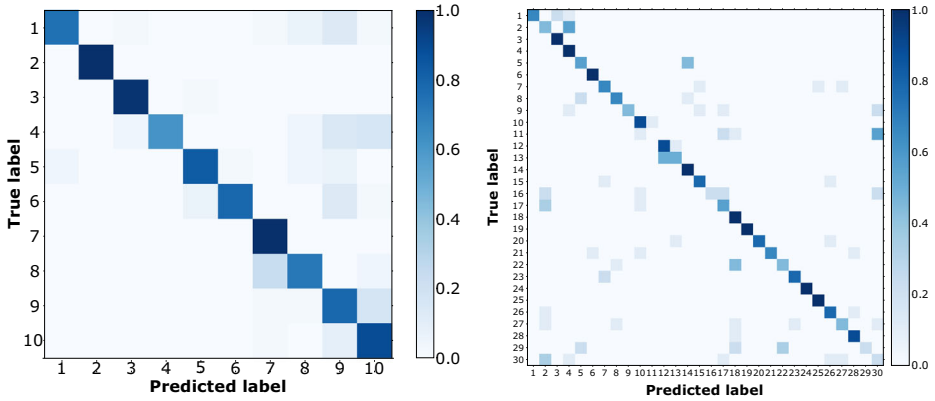


Fig. 10 Confusion matrix of test results for the 3DFCNN on NTU RGB+D with 60 human actions (CV evaluation). Action indexes in figure correspond to the indexes used in author's webpage [60]

Table 3 Top 10 accurate actions and confused pairs for the proposed model, including accuracy recognition per action

Top 10 recognized actions		Top 10 confused actions*	
1) Stand up	(97.47%)	1) Reading → Writing	(39.56%)
2) Jump up	(97.15%)	2) Writing → Type on a keyboard	(40.82%)
3) Hopping	(96.84%)	3) Play with phone/tablet → Type on a keyboard	(44.62%)
4) Shake head	(96.84%)	4)Sneeze/cough → Nausea/vomiting	(53.80%)
5) Falling down	(96.84%)	5) Phone call → Eat meal	(54.43%)
6) Sit down	(96.52%)	6) Take off a shoe → Put on a shoe	(56.96%)
7) Take off a hat/cap	(95.87%)	7) Headache → Brush teeth	(61.39%)
8) Walking apart	(95.57%)	8) Point to something → Taking a selfie	(61.71%)
9) Hugging	(95.25%)	9) Eat meal → Phone call	(62.66%)
10) Cheer up	(94.94%)	10) Brush teeth → Drink water	(64.87%)
Average accuracy = 96.33%		Average accuracy = 54.08%	

*Numbers between parenthesis are the recognition accuracy of true action (before the arrow)



(a) NWUCLA dataset (b) UWA3DII dataset

Fig. 11 Confusion matrix of test results for the 3DFCNN on NWUCLA and UWA3DII datasets

The confusion matrices have also been obtained for NWUCLA and UWA3DII datasets which, as it has been explained, include a lower number of actions. The results are shown in Fig. 11.

As it can be seen, 3DFCNN is able to recognize action with a high accuracy for both datasets.

5.3 Comparison with state-of-the-art methods

The three datasets described above have also been used to compare the method proposed in this work with other alternatives in the state of the art. Firstly, the 3DFCNN has been trained and evaluated with the large scale NTU RGB+D dataset, widely used by deep learning based methods due to the high number of samples available. The obtained results and the comparison against other state-of-the-art proposals is exposed in Section 5.3.1.

To test the generalization capacity of the network, as well as its ability to adapt to other environments, it has been evaluated in two other image datasets that include depth information: NWUCLA and UWA3DII. These are two datasets with a much lower number of samples, which makes the training of DNNs difficult (so they are mainly used for the evaluation of proposals based on non-DNN-based techniques). Furthermore, the used sensor is the Kinect V1, whose measurements contain more noise than the Kinect V2 camera, due to its different depth acquisition technology. The results obtained with NWUCLA and UWA3DII are shown in Section 5.3.2. Since the number of samples is too small for most of the deep learning based approaches, the results in this section are compared to several state-of-the-art methods based on non-DNN-based techniques.

5.3.1 Comparison with NTU RGB+D dataset

Table 4 presents some results obtained by different proposals in the literature for the NTU RGB+D dataset.

The results are divided into two different sections according to the modality used: first, results obtained by methods that use 3D skeletons, and second, approaches that only use depth data (like the proposal in this paper). It is worth highlighting that all the methods

Table 4 Total average accuracy (%) from different methods on NTU RGB+D dataset

Method	CS	CV	Preprocessing
Modality: 3D Skeleton			
ST-LSTM + Trust Gate (2016) [43]	69.2	77.7	Tree-like graph
Clips + CNN + MTLN (2017) [26]	79.57	84.83	Encoding clips
AGC-LSTM (2019) [61]	**89.2**	**95.0**	Feature processing
Modality: Depth			
DDI+DDNI+DDMNI (2018) [78]	**87.08**	84.22	Dynamic images
HDDPDI (2019) [88]	82.43	**87.56**	Dynamic images
MVDI+CNN (2019) [89]	84.6	87.3	Dynamic images
Proposed method (3DFCNN)	78.13	80.37	–

included in Table 4 are based on DNNs. It also includes a column with a brief description of the preprocessing that is used by each method. The proposed method, 3DFCNN, achieves an accuracy of 78.13% (CS) and 80.37% (CV) on NTU RGB+D dataset. As shown in Table 4, although the 3DFCNN model does not overcome the state-of-the-art methods like, for instance, the one proposed in [89], it gets a remarkable and competitive recognition accuracy considering the challenging dataset, NTU RGB+D.

Moreover, the performance of the proposed method gets more valuable when it is taken into account its simplicity and the absence of a costly preprocessing, as shown in the last column of Table 4, being the only work that does not make a big conversion of the input (for example, turning the video sequence into dynamic images like the rest of depth-based works do). This allows the method to be considerably fast, which is a great advantage when it is compared with the complex methods from the state-of-the-art. Table 5 shows an estimation of the computational cost in terms of average time consumption per video in test phase as reported by some previous methods.

The first four computational cost results are taken from [78], and, although they were computed with a different hardware than the present work, they give an idea of the reduced computational cost of our method when it is compared, for instance, with the method that uses three sets of dynamic images (DDI+DDNI+DDMNI [78]), which achieves some of the highest recognition rate on NTU RGB+D dataset, see Table 4. For the sake of clarity, these four results were computed on a gesture RGB-D dataset, and only the last one was also used

Table 5 Time consumption comparison of some action recognition methods with available data. The time values are the average of several time consumption values from different video samples of the dataset. See the text for more details

Method	Time (s)	Dataset
MSFK+DeepID [71]	41.00	CGD, CAD-60, MSR Daily Activity 3D
SFAM [82]	6.33	ChaLearn LAP IsoGD, M^2I
WHDMM [81]	1.04	MSR-Action3DExt, UTKinect-Action
DDI+DDNI+DDMNI [78]	62.03	NTU RGB+D, ChaLearn LAP IsoGD
MVDI+CNN [89]	51.02	NTU RGB+D, UWA3DII, NWUCLA
Proposed method (3DFCNN)	**1.09**	NTU RGB+D, UWA3DII, NWUCLA

Table 6 Recognition accuracy (%) for different depth maps-based methods (CCD [7], HON4D [51], HOPC [53] and MVDI [89]) on the UWA3DII dataset. The evaluation criteria for this dataset consists of using two camera views for training and the other two for testing

Train	V_1+V_2		V_1+V_3		V_1+V_4		V_2+V_3		V_2+V_4		V_3+V_4		Mean
Test	V_3	V_4	V_2	V_4	V_2	V_3	V_1	V_4	V_1	V_3	V_1	V_2	
CCD	10.5	13.6	10.3	12.8	11.1	8.3	10.0	7.7	13.1	13.0	12.9	10.8	11.2
HON4D	31.1	23.0	21.9	10.0	36.6	32.6	47.0	22.7	36.6	16.5	41.4	26.8	28.9
HOPC	52.7	51.8	59.0	**57.5**	42.8	44.2	58.1	38.4	63.2	43.8	66.3	48.0	52.2
MVDI	**77.0**	**59.5**	**68.3**	57.2	57.8	**72.9**	**80.3**	**51.3**	**76.6**	**69.5**	**78.8**	67.9	68.1
3DFCNN	68.3	54.6	66.7	51.5	**68.2**	67.2	74.3	49.7	75.1	61.9	73.5	**88.3**	66.6

on the action recognition dataset used in this paper, as it can be seen in the last column of Table 5.

The 3DFCNN model presented in this paper achieves therefore the lowest time consumption among the works on NTU RGB+D dataset, with an average time per 30-frames video sequence of 1.09 s, in which around the 90% is due to the preprocessing steps of background removal and image cropping. This value has been computed using a NVIDIA GeForce GTX 1080 with 8 GB and an Intel(R) Core(R) i7-7700 CPU at 3.60 GHz, through a set of 10 000 randomly-chosen depth video samples from the NTU RGB+D dataset. The same GPU is used in [89], where multi-view dynamic images (MVDI) were used, together with an Intel(R) Xeon(R) E5-2630 V3 CPU running at 2.4 GHz. It can be considered a similar workstation than the one used for this paper, thus allowing a fair comparison of computational costs. In that paper, they used the CPU to generate the multi-view dynamic images, which was the main source of computational cost, yielding a total average time consumption of 51.02 s per video. Therefore, this high computational cost prevent this method to be used in real time applications like, e.g., video surveillance and health care. However, the 3DFCNN method gets, at least, an average time consumption that is one order of magnitude lower. In summary, the method proposed in the present paper attains both a sufficient high action recognition accuracy and yet a very low computational cost.

5.3.2 Comparison with UWA3DII and NWUCLA datasets

Tables 6 and 7 show the results obtained with UWA3DII and NWUCLA datasets, respectively.

The results with both datasets are shown in the same section because of the similarities between their characteristics (type of sensor, number of people, number of samples, etc.). There is also presented the comparison to other three different proposals. It should be noted

Table 7 Recognition accuracy (%) for different depth maps-based methods on the Northwestern-UCLA dataset. The evaluation criteria for this dataset consists of using the first two camera views for training and the remaining view for testing

Method	Acc.
CCD [7]	34.4
HON4D [51]	39.9
HOPC [53]	80.0
MVDI [89]	**84.2**
Proposed method (3DFCNN)	83.6

that most of the previous works that use these datasets are based in non-DNN-based methods, since the reduced number of samples make it difficult training a DNN. On one hand, the first three works propose different new descriptors for action recognition: the Comparative Coding Descriptor (CCD) [7], the histogram of the surface normal orientation in the 4D space (HON4D) [51], and the Histogram of Oriented Principal Components (HOPC) [54]. On the other hand, the proposal in [89] is based on a CNN combined to dynamic images, but it requires a costly preprocessing to obtain the dynamic images.

As it can be seen in Tables 6 and 7, despite the small number of samples for fine-tuning in these datasets, the results of the 3DFCNN proposed in this paper outperforms the approaches based on non-DNN-based methods. Besides, the results are close to those obtained using the proposal in [89], surpassing them in some cases. These results are even more significant considering the reduced computational cost of the proposal compared to [89].

6 Conclusions

This paper proposes the 3DFCNN, an end-to-end trainable deep learning approach for HAR from depth videos. The model is a 3D fully convolutional neural network, which automatically encodes spatial and temporal patterns of depth sequences without a costly preprocessing. Furthermore, an efficient data generation system and a particular training strategy were proposed. Newly appeared deep learning techniques like learning rate range test, cyclical learning rate schedule and fully convolutional architecture were used in order to improve the model performance. An exhaustive experimental evaluation of the proposal has been carried out, using three different publicly available datasets. Experimental results on the large-scale NTU RGB+D dataset show the proposed method achieves action recognition accuracy close to state-of-the-art deep learning based methods, while drastically reducing the computational cost because of its relatively simple structure and scarce preprocessing. This property would allow the proposed 3DFCNN model to run on real time applications, like video surveillance, health care services, video analysis and human-computer interaction. Besides, results within smaller NWUCLA and UWA3DII datasets show that the proposal reliability overtakes that of different methods based on non-DNN-based computer vision techniques, and obtains results comparable to those from other state-of-the-art methods based on deep learning.

As most of the action recognition methods, the 3DFCNN model tends to confuse similar actions in which there are just small discriminatory objects or short motions, like within actions *writing* and *reading*. Improving recognition accuracy of such actions is still an open problem and constitutes a line of future work for the proposal here presented.

Acknowledgements Portions of the research in this paper used the "NTU RGB+D (or NTU RGB+D 120) Action Recognition Dataset" made available by the ROSE Lab at the Nanyang Technological University, Singapore.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Al-Akam R, Paulus D, Gharabaghi D (2018) Human action recognition based on 3d convolution neural networks from rgbd videos. In: WSCG 2018: Poster papers proceedings: 26th international conference in central europe on computer graphics, visualization and computer vision, pp 18–26
2. Ashraf N, Sun C, Foroosh H (2014) View invariant action recognition using projective depth. Comput Vis Image Underst 123:41–52
3. Baptista-Ríos M, Martínez-García C, Losada-Gutiérrez C, Marrón-Romera M (2016) Human activity monitoring for falling detection. a realistic framework. In: 2016 International conference on indoor positioning and indoor navigation (IPIN), pp 1–7. https://doi.org/10.1109/IPIN.2016.7743617
4. Chaquet JM, Carmona EJ, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. Comput Vis Imag Underst 117(6):633–659. https://doi.org/10.1016/j.cviu.2013.01.013. http://www.sciencedirect.com/science/article/pii/S1077314213000295
5. Chen C, Jafari R, Kehtarnavaz N (2017) A survey of depth and inertial sensor fusion for human action recognition. Multimed Tools Appl 76(3):4405–4425
6. Chen C, Liu K, Kehtarnavaz N (2016) Real-time human action recognition based on depth motion maps. J Real-Time Imag Process 12(1):155–163
7. Cheng Z, Qin L, Ye Y, Huang Q, Tian Q (2012) Human daily action analysis with multi-view and color-depth data. In: European conference on computer vision. Springer, pp 52–61
8. Chou KP, Prasad M, Wu D, Sharma N, Li DL, Lin YF, Blumenstein M, Lin WC, Lin CT (2018) Robust feature-based automated multi-view human action recognition system. IEEE Access 6:15283–15296
9. Das S, Thonnat M, Sakhalkar K, Koperski M, Bremond F, Francesca GKompatsiaris I, Huet B, Mezaris V, Gurrin C, Cheng WH, Vrochidis S (eds) (2019) A new hybrid architecture for human activity recognition from rgb-d videos. Springer International Publishing, Cham
10. Dawar N, Chen C, Jafari R, Kehtarnavaz N (2017) Real-time continuous action detection and recognition using depth images and inertial signals. In: 2017 IEEE 26th international symposium on industrial electronics (ISIE). IEEE, pp 1342–1347
11. Dipakkr (2018) 3d-cnn action recognition. https://github.com/dipakkr/3d-cnn-action-recognition
12. Farooq A, Won CS (2015) A survey of human action recognition approaches that use an rgb-d sensor. IEIE transactions on smart processing & computing 4(4):281–290
13. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1933–1941
14. Gebert P, Roitberg A, Haurilet M, Stiefelhagen R (2019) End-to-end prediction of driver intention using 3d convolutional neural networks. In: 2019 IEEE Intelligent vehicles symposium (IV), pp 969–974. https://doi.org/10.1109/IVS.2019.8814249
15. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J (2016) Lstm: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems 28(10):2222–2232
16. Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans Cybern 43(5):1318–1334
17. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition
18. Hou Y, Wang S, Wang P, Gao Z, Li W (2017) Spatially and temporally structured global to local aggregation of dynamic depth information for action recognition. IEEE Access 6:2206–2219
19. Hsu Y. P., Liu C., Chen T. Y., Fu L. C. (2016) Online view-invariant human action recognition using rgb-d spatio-temporal matrix. Pattern Recogn 60:215–226. https://doi.org/10.1016/j.patcog.2016.05.010. http://www.sciencedirect.com/science/article/pii/S0031320316300930
20. Hu JF, Zheng WS, Lai J, Zhang J (2015) Jointly learning heterogeneous features for rgb-d activity recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR)

21. Hu JF, Zheng WS, Pan J, Lai J, Zhang J (2018) Deep bilinear learning for rgb-d action recognition. In: The european conference on computer vision (ECCV)
22. Hu X, Yang K, Fei L, Wang K (2019) Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE International conference on image processing (ICIP), pp 1440–1444. https://doi.org/10.1109/ICIP.2019.8803025
23. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167
24. Ji S, Xu W, Yang M, Yu K (2012) 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1):221–231
25. Jozefowicz R, Zaremba W, Sutskever I (2015) An empirical exploration of recurrent network architectures. In: International conference on machine learning, pp 2342–2350
26. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3288–3297
27. Khaire P, Kumar P, Imran J (2018) Combining cnn streams of rgb-d and skeletal data for human activity recognition. Pattern Recogn Lett 115:107–116. https://doi.org/10.1016/j.patrec.2018.04.035. http://www.sciencedirect.com/science/article/pii/S0167865518301636. Multimodal Fusion for Pattern Recognition
28. Khurana R, Kushwaha AKS (2018) Deep learning approaches for human activity recognition in video surveillance-a survey. In: 2018 First international conference on secure cyber computing and communication (ICSCCC). IEEE, pp 542–544
29. Kim TS, Reiter A (2017) Interpretable 3d human action analysis with temporal convolutional networks. In: 2017 IEEE Conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 1623–1631
30. Kingma D. P., Ba J. (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
31. Ko K. E., Sim K. B. (2018) Deep convolutional framework for abnormal behavior detection in a smart surveillance system. Eng Appl Artif Intell 67:226–234. https://doi.org/10.1016/j.engappai.2017.10.001. http://www.sciencedirect.com/science/article/pii/S0952197617302579
32. Kong J., Liu T., Jiang M. (2019) Collaborative multimodal feature learning for rgb-d action recognition. J Visual Commun Imag Represent 59:537–549. https://doi.org/10.1016/j.jvcir.2019.02.013. http://www.sciencedirect.com/science/article/pii/S104732031930063X
33. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25. Curran Associates Inc, pp 1097–1105
34. Lange R, Seitz P (2001) Solid-state time-of-flight range camera. IEEE J Quantum Electron 37(3):390–397. https://doi.org/10.1109/3.910448
35. Laraba S, Brahimi M, Tilmanne J, Dutoit T (2017) 3d skeleton-based action recognition by representing motion capture sequences as 2d-rgb images. Computer Animation and Virtual Worlds 28(3-4):e1782
36. Li C, Zhong Q, Xie D, Pu S (2017) Skeleton-based action recognition with convolutional neural networks. In: 2017 IEEE International conference on multimedia & expo workshops (ICMEW). IEEE, pp 597–600
37. Li S, Li W, Cook C, Zhu C, Gao Y (2018) Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5457–5466
38. Li X, Hou Z, Liang J, Chen C (2020) Human action recognition based on 3d body mask and depth spatial-temporal maps. Multimed Tools Appl 79(47):35761–35778
39. Liu A. A., Nie W. Z., Su Y. T., Ma L., Hao T., Yang Z. X. (2015) Coupled hidden conditional random fields for rgb-d human action recognition. Signal Processing 112:74–82. https://doi.org/10.1016/j.sigpro.2014.08.038. http://www.sciencedirect.com/science/article/pii/S0165168414004022. Signal Processing and Learning Methods for 3D Semantic Analysis
40. Liu B, Cai H, Ju Z, Liu H (2019) Rgb-d sensing based human action and interaction analysis: a survey. Pattern Recogn 94:1–12
41. Liu J, Akhtar N, Ajmal M (2018) Viewpoint invariant action recognition using rgb-d videos. IEEE Access 6:70061–70071
42. Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC (2019) Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2019.2916873
43. Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal lstm with trust gates for 3d human action recognition. In: European conference on computer vision. Springer, pp 816–833

44. Liu K, Liu W, Gan C, Tan M, Ma H (2018) T-c3d: Temporal convolutional 3d network for real-time action recognition. In: Thirty-second AAAI conference on artificial intelligence
45. Liu Z, Gao G, Qin AK, Wu T, Liu CH (2019) Action recognition with bootstrapping based long-range temporal context attention. In: Proceedings of the 27th ACM International Conference on Multimedia, pp 583–591
46. Luo Z, Peng B, Huang DA, Alahi A, Fei-Fei L (2017) Unsupervised learning of long-term motion dynamics for videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2203–2212
47. Maas AL, Hannun AY, Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml, vol 30, p 3
48. Martinez M., Yang K., Constantinescu A., Stiefelhagen R. (2020) Helping the blind to get through covid-19: Social distancing assistant using real-time semantic segmentation on rgb-d video. Sensors 20(18). https://doi.org/10.3390/s20185202. https://www.mdpi.com/1424-8220/20/18/5202
49. Ning X, Duan P, Li W, Shi Y, Li S (2020) A cpu real-time face alignment for mobile platform. IEEE Access 8:8834–8843. https://doi.org/10.1109/ACCESS.2020.2964838
50. Ning X, Xu S, Li W, Nie S (2020) Fegan: Flexible and efficient face editing with pre-trained generator. IEEE Access 8:65340–65350. https://doi.org/10.1109/ACCESS.2020.2985086
51. Oreifej O, Liu Z (2013) Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 716–723
52. Poppe R (2010) A survey on vision-based human action recognition. Image and Vision Computing 28(6):976–990
53. Rahmani H, Mahmood A, Huynh D, Mian A (2016) Histogram of oriented principal components for cross-view action recognition. IEEE Trans Pattern Anal Mach Intell 38(12):2430–2443. https://doi.org/10.1109/TPAMI.2016.2533389
54. Rahmani H, Mahmood A, Huynh DQ, Mian A (2014) Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In: European conference on computer vision. Springer, pp 742–757
55. Sadanand S, Corso JJ (2012) Action bank: a high-level representation of activity in video. In: Computer vision and pattern recognition (CVPR), 2012 IEEE conference on. IEEE, pp 1234–1241
56. Sarfraz MS, Murray N, Sharma V, Diba A, Van Gool L, Stiefelhagen R (2021) Temporally-weighted hierarchical clustering for unsupervised action segmentation. arXiv:2103.11264
57. Schindler K, Van Gool L (2008) Action snippets: How many frames does human action recognition require? In: 2008 IEEE Conference on computer vision and pattern recognition. IEEE, pp 1–8
58. Sell J, O'Connor P (2014) The Xbox one system on a chip and Kinect sensor. Micro, IEEE 34(2):44–53. https://doi.org/10.1109/MM.2014.9
59. Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+d: a large scale dataset for 3d human activity analysis. In: The IEEE conference on computer vision and pattern recognition (CVPR)
60. Shahroudy A., Liu J., Ng T. T., Wang G. (2016) NTU RGB+D Action Recognition dataset. Available online: http://rose1.ntu.edu.sg/datasets/actionrecognition.asp (Last access 12/11/2019)
61. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1227–1236
62. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems 27. Curran Associates Inc, pp 568–576
63. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition
64. Simonyan K., Zisserman A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
65. Singh T, Vishwakarma DK (2019) Human activity recognition in video benchmarks: a survey. In: Rawat BS, Trivedi A, Manhas S, Karwal V (eds) Advances in signal processing and communication. Springer, Singapore, pp 247–259
66. Siyal MR, Ebrahim M, Adil SH, Raza K (2020) Human action recognition using convlstm with gan and transfer learning. In: 2020 International conference on computational intelligence (ICCI), pp 311–316. https://doi.org/10.1109/ICCI51257.2020.9247670
67. Smith LN (2017) Cyclical learning rates for training neural networks. In: 2017 IEEE Winter conference on applications of computer vision (WACV), pp 464–472. IEEE
68. Song S, Lan C, Xing J, Zeng W, Liu J (2017) An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Thirty-first AAAI conference on artificial intelligence

69. Tafazzoli F., Safabakhsh R. (2010) Model-based human gait recognition using leg and arm movements. Engineering Applications of Artificial Intelligence 23(8):1237–1246. https://doi.org/10.1016/j.engappai.2010.07.004. http://www.sciencedirect.com/science/article/pii/S0952197610001417
70. Tian D, Lu ZM, Chen X, Ma LH (2020) An attentional spatial temporal graph convolutional network with co-occurrence feature learning for action recognition. Multimed Tools Appl, 1–19
71. Wan J, Guo G, Li SZ (2015) Explore efficient local features from rgb-d data for one-shot learning gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(8):1626–1639
72. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: The IEEE international conference on computer vision (ICCV)
73. Wang J, Nie X, Xia Y, Wu Y, Zhu S (2014) Cross-view action modeling, learning, and recognition. In: 2014 IEEE Conference on computer vision and pattern recognition, pp 2649–2656. https://doi.org/10.1109/CVPR.2014.339
74. Wang L, Ding Z, Tao Z, Liu Y, Fu Y (2019) Generative multi-view human action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp 6212–6221
75. Wang L., Huynh D. Q., Koniusz P. (2020) A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing 29:15–28. https://doi.org/10.1109/tip.2019.2925285
76. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. Springer, pp 20–36
77. Wang L, Xu Y, Cheng J, Xia H, Yin J, Wu J (2018) Human action recognition by learning spatio-temporal features with deep neural networks. IEEE Access 6:17913–17922
78. Wang P, Li W, Gao Z, Tang C, Ogunbona PO (2018) Depth pooling based large-scale 3-d action recognition with convolutional neural networks. IEEE Transactions on Multimedia 20(5):1051–1061
79. Wang P, Li W, Gao Z, Tang C, Zhang J, Ogunbona P (2015) Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring. In: Proceedings of the 23rd ACM international conference on Multimedia. ACM, pp 1119–1122
80. Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona P (2015) Deep convolutional neural networks for action recognition using depth map sequences. arXiv:1501.04686
81. Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona PO (2015) Action recognition from depth maps using deep convolutional neural networks. IEEE Transactions on Human-Machine Systems 46(4):498–509
82. Wang P, Li W, Gao Z, Zhang Y, Tang C, Ogunbona P (2017) Scene flow to action map: a new representation for rgb-d based action recognition with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 595–604
83. Wang P, Li W, Liu S, Gao Z, Tang C, Ogunbona P (2016) Large-scale isolated gesture recognition using convolutional neural networks. In: 2016 23Rd international conference on pattern recognition (ICPR). IEEE, pp 7–12
84. Wang P., Li W., Ogunbona P., Wan J., Escalera S. (2018) Rgb-d-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding 171:118–139. https://doi.org/10.1016/j.cviu.2018.04.007. http://www.sciencedirect.com/science/article/pii/S1077314218300663
85. Wang P, Wang S, Gao Z, Hou Y, Li W (2017) Structured images for rgb-d action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 1005–1014
86. Weinland D, Ronfard R, Boyer E (2011) A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding 115(2):224–241
87. Weng J, Weng C, Yuan J (2017) Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4171–4180
88. Wu H., Ma X., Li Y. (2019) Hierarchical dynamic depth projected difference images–based action recognition in videos with convolutional neural networks. International Journal of Advanced Robotic Systems 16(1):1729881418825093
89. Xiao Y, Chen J, Wang Y, Cao Z, Zhou JT, Bai X (2019) Action recognition for depth video using multi-view dynamic images. Inf Sci 480:287–304
90. Xu B, Wang N, Chen T, Li M (2015) Empirical evaluation of rectified activations in convolutional network. arXiv:1505.00853
91. Xu Y, Hou Z, Liang J, Chen C, Jia L, Song Y (2019) Action recognition using weighted fusion of depth images and skeleton's key frames. Multimed Tools Appl 78(17):25063–25078

92. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence
93. Yang K, Zhang J (2021) Reiß, S., Hu, X., Stiefelhagen R.: Capturing omni-range context for omnidirectional segmentation
94. Zhang B, Wang L, Wang Z, Qiao Y, Wang H (2016) Real-time action recognition with enhanced motion vector cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2718–2726
95. Zhang HB, Zhang YX, Zhong B, Lei Q, Yang L, Du JX, Chen DS (2019) A comprehensive survey of vision-based human action recognition methods. Sensors 19(5):1005
96. Zhang J, Han Y, Tang J, Hu Q, Jiang J (2017) Semi-supervised image-to-video adaptation for video action recognition. IEEE transactions on cybernetics 47(4):960–973
97. Zhang J, Li W, Ogunbona PO, Wang P, Tang C (2016) Rgb-d-based action recognition datasets: A survey. Pattern Recognition 60:86–105. https://doi.org/10.1016/j.patcog.2016.05.019. http://www.sciencedirect.com/science/article/pii/S0031320316301029
98. Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N (2017) View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2117–2126
99. Zhang Z (2012) Microsoft kinect sensor and its effect. IEEE multimedia 19(2):4–10
100. Zhao Y, Liu Z, Yang L, Cheng H (2012) Combing rgb and depth map features for human activity recognition. In: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp 1–4

Affiliations

Adrián Sánchez-Caballero[1] · Sergio de López-Diz[1] · David Fuentes-Jimenez[1] · Cristina Losada-Gutiérrez[1] · Marta Marrón-Romera[1] · David Casillas-Pérez[2] · Mohammad Ibrahim Sarker[1]

Adrián Sánchez-Caballero
adrian.sanchez@uah.es

Sergio de López-Diz
s.lopezd@edu.uah.es

David Fuentes-Jimenez
d.fuentes@edu.uah.es

Marta Marrón-Romera
marta.marron@uah.es

David Casillas-Pérez
david.casillas@urjc.es

Mohammad Ibrahim Sarker
ibrahim.sarker@uah.es

[1] Department of Electronics, University of Alcalá, Ctra, Madrid-Barcelona, km. 33600, 28805, Alcalá de Henares, Spain

[2] Department of Signal Processing and Communications, Universidad Rey Juan Carlos, Fuenlabrada, Madrid, Spain