The Visual Computer (2022) 38:993–1013
https://doi.org/10.1007/s00371-021-02064-y

ORIGINAL ARTICLE

DTR-HAR: deep temporal residual representation for human activity recognition

Hend Basly[1] · Wael Ouarda[2] · Fatma Ezahra Sayadi[3] · Bouraoui Ouni[1] · Adel M. Alimi[2]

Accepted: 5 January 2021 / Published online: 15 February 2021

Abstract

Human activity recognition (HAR) is a highly prized application in the pattern recognition and the computer vision fields. Up till now, deep neural networks have acquired big attention in computer studies and image processing fields, and have generated significant results. In this paper, we propose a deep temporal residual system for daily living activity recognition that aims to enhance spatiotemporal feature representation in order to improve the HAR system performance. To this end, we adopt a deep residual convolutional neural network (RCN) to retain discriminative visual features relayed to appearance and long short-term memory neural network to capture the long-term temporal evolution of actions. The latter was considered to implement time dependencies occurring when carrying out the activity to enhance features extracted from the RCN network by adding time information to address the dynamic activity recognition problem as a sequence labeling job. The deep temporal residual model for human activity recognition system is performed on two benchmark publicly available datasets: MSRDailyActivity3D and CAD-60. the proposed system achieves very competitive results when compared to others from the state of the art.

Keywords Daily living activity recognition · Convolutional neural network (CNN) · Long short-term memory (LSTM) · Video surveillance

1 Introduction

Human activity recognition in video sequences has attracted significant attention in computer vision in recent decades, because of their very broad scope such as home care or daily living assistance for the elderly/disabled. Research progress in this field has been faster than expected. Our research proposes a daily living activity recognition that aims to observe persons and automatically identify what they are doing as actions in a video sequence. For human action classification, when we deal with video sequences and not just an image, an action can be defined as the analyze of what a person is doing in front of the camera. So, it describes the main events in the video and it can be represented by a sequence of frames which people can understand seamlessly by using reading contents of sequential frames. The temporal dynamics between frames in a video sequence provide more motion information to recognize the action. Here, the task becomes more computationally expensive because video sequences are composed of hundreds of frames that require to be executed separately. In complex video scenes, visual data need to be analyzed and transformed into a format that represents the visual content

✉ Hend Basly
basly.hend@gmail.com

Wael Ouarda
wael.ouarda@ieee.org

Fatma Ezahra Sayadi
sayadi_fatma@yahoo.fr

Bouraoui Ouni
ouni_bouraoui@yahoo.fr

Adel M. Alimi
adel.alimi@ieee.org

1 NOCCS–Lab.: Networked Objects Control and Communication Systems Laboratory, National Engineering School of Sousse (ENISO), University of Sousse, BP 264, 4023 Erriadh, Sousse, Tunisia

2 REGIM–Lab.: REsearch Groups in Intelligent Machines, National Engineering School of Sfax (ENIS), University of Sfax, BP 1173, 3038 Sfax, Tunisia

3 EμE-Lab.: Electronics and Microelectronics Laboratory, Faculty of Sciences of Monastir (FSM), University of Monastir, Environment Avenue, 5019 Monastir, Tunisia

effectively. For this purpose, human parsing techniques [1–4] can be adopted to reduce potentially the searching space and to provide contextual hints for human action recognition, in order to obtain the high-level semantic for an image. Several human activity recognition (HAR) approaches have been proposed, ranging from traditional method-based handcrafted feature descriptor to deep learning-based methods. HAR approaches based on deep learning have focused great researcher attention by the considerable improvement in the recognition accuracy since they overcome several obstacles confronted by traditional handcrafted feature-based methods. The strength of deep neural networks is provided by their ability to act as automatic feature extractor to represent hidden patterns by backpropagation and learn long-term temporal dependencies. Handcrafted feature-based methods extract low-level features from video data and feed them to a classifier, such as K-nearest neighbor (KNN), hidden Markov model (HMM), support vector machine (SVM) to recognize actions. These approaches [5,6] are limited by their difficulty to process lengthy videos containing sequential and continuous information. They are ineffective to recognize more than one-person action in a scene such as space–time–volume (STV)-based methods. Some used techniques such as space–time features (STF) and motion history image (MHI) are more appropriate to simple datasets, but with hybrid methods combining different feature extractors (such as HOG, SURF and SIFT), systems become more expensive in terms of computational complexity. The action recognition is based on action detection taking into account a sequence of frames to determine the overall action. The action detection can likewise be used to describe the dominant events in a video. The human activity recognition task can be described by taking sequential data as input to provide one single classification as output. The video representation process has to be simple and efficient to enforce by selecting and extracting powerful and discriminant high-level features, which are robust to viewpoints and appearance changes. The good representation of spatiotemporal features from video frame sequences is critical to build and train a robust model serving to accurately recognize the activity. Convolutional neural networks (CNNs) are renowned for their great potential in identifying implicit features for visual appearance data by backpropagation, so they act as automatic feature extractor without any artificial intervention. Nevertheless, they are unable to represent long-term temporal dynamics between frames of the entire video. Recurrent neural networks (RNN) are suitable to extract hidden patterns of data in both space and temporal domains. To reduce the impact of vanishing gradient problem, LSTM has been elaborated to handle time series data and to acquire long-term dependencies between frames. Since we are facing the daily living activity recognition problem, we have to be very quick to make prompt decision. A model that should make an accurate prompt decision should learn from scratch. More recently, He et al. [7] have proposed deep residual networks which are very helpful to attenuate the degradation problem when training deeper networks, since they involve more than hundreds of layers by announcing a novel architecture which includes residual connections making the residual networks easier to optimize than the plain networks. The residual connections facilitate the learning of the identity function; therefore, they facilitate the propagation of the gradient from the outlet to the network entrance to enable its efficient training. Before the development of this type of network architecture, it was impossible to train a network with more than 25 layers. As the layers got deeper and the gradients got smaller, performance inevitably deteriorated; the error was no longer propagated correctly and the update of the weights was directly affected. Residual neural networks have made it possible to go beyond this limitation. Their architecture allows the creation of very deep neural networks, with better accuracy than those with linear architectures because they allow to extract more information and thus have a more advanced analysis of images. Since in this research work we use databases containing many videos which are in turn split into a large number of images, learning the network therefore requires a fairly large amount of data. Aware of this problem, we explore the daily living activity recognition issue by proposing a deep temporal residual model for human activity recognition using raw color (RGB) data information. Our proposed DTR-HAR model converges faster than previous deep neural networks with linear architecture (AlexNet [8], GoogLeNet [9], VGGNet [10]). In fact, this model was based on both the skill of the convolutional neural network to extract spatial features implicitly and the power of recurrent neural networks (LSTM) to handle with time series data in order to model the temporal evolution of the activity. More particularly, inspired by the recent advance of residual neural networks [7] and video classification [11], we develop a deep temporal residual network composed of two subnetworks: The first is a deep residual convolutional neural network (RCN) with dense residual crossconnections between the layers to encode sequential data and extract visual appearance features using a pre-trained CNN architecture based on ImageNet. To enhance features extracted from the RCN network a second subnetwork is developed which is an LSTM model that takes advantage of capturing the long-term temporal evolution of actions by sequencing the learning features. To simplify the task, we separate the video into many sequence video frames, which will be used as input to our model. The residual convolutional stream-based pre-trained ResNet-101 [7] model encodes the video frames as residual spatial features. The latter are enhanced by applying activations for the last fully connected layer of the trained RCN network to generate a new encoded feature representation. These features are fed in second step, into the LSTM model to extract temporal features by sequencing

them taking into account the temporal dependencies of activities in the frame sequences and to finally output the activity class of each video.

Our paper has the following main contributions. We first propose a deep temporal residual model for human activity recognition (DTR-HAR) using raw color (RGB) data information by combining two deep learning architectures: A feed forward neural network with end-to-end structure-based residual CNN architecture is carried out to extract visual residual spatial features relayed to the appearance of persons automatically and a LSTM network to capture the long-term temporal evolution of actions in video streaming. CNN and LSTM networks are fused together to generate new encoded feature representations useful for automatic activity recognition. Then, we have reduced the required training time and avoid our classifier from overfitting by ensuring the best weight initialization, given the quite small number of available datasets. We used transfer learning to fine-tune the parameters of a pre-trained architecture that was trained on ImageNet dataset. Finally, we expand our datasets artificially by applying some transformations and geometric deformations as data augmentation technique to prevent the model from overfitting and to better discriminate between pertinent characteristics. Through considerable experimentations, we have spotlighted the benefits of our framework by validating it on two publicly available datasets, i.e., MSR-DailyActivity3D dataset and CAD-60. The obtained results are competitive to the state-of-the-art performance (91.65% on MSRDailyActivity3D and 91.18% on CAD-60) proving the efficiency and the utility of the proposed approach.

The organization of the remainder of this article is as follows: In Sect. 2, we highlight the related works; in Sect. 3, we elaborate our methodology in detail. We validate our experimentations and discuss our results in Sect. 4. Finally, we conclude our paper in Sect. 5.

2 Related works

This section reviews the most relevant existent methods from literature related to action recognition, beginning from handcraft-based representation approaches until deep learning-based ones.

2.1 Handcraft-based approaches

In action recognition, classical descriptors used to extract video representations/features are extended to include temporal dimension [12–15]. The scale-invariant feature transform (SIFT) in [16] was extended to 3D-SIFT [12] and used to encode spatiotemporal local information to bring more robustness to the system. The speeded up robust feature (SURF) technique, which analyses each input frame at different scales to ensure its invariance to scale changes by applying a scale-invariant descriptor and rotation, was extended to 3D SURF [17,18]. In [13], the authors developed the concepts of histogram of oriented gradient (HoG) descriptor in video sequences to obtain HoG-3D. The authors extended images to videos to get 3D gradient vectors. Subsequently, descriptor parameters were calculated and optimized for action recognition. Reference [19] used the motion history image (MHI) to represent the motion direction, the foreground image (FI) to get the background subtraction and the HOG descriptor to characterize the magnitude and direction of corners and edges. These three types of feature representation were then merged and classified using a simulated annealing multiple instance learning support vector machine (SMILE-SVM). Furthermore, for data modeling, several works have relied on a sparse representation to extract the key characteristics used by the human activity recognition systems [20–22]. In [20], the authors presented a human daily activity recognition framework based on the selection of an overcomplete dictionary to construct sparse representations using signals sampled from wearable sensors. In [21], Bhattacharya et al. introduced a sparse-coding technique to represent sensor data for an unsupervised estimation using a codebook of basis vectors that includes characteristic and latent movement patterns for human activity recognition. Also, the work of [22] proposed a multi-temporal dictionary learning strategy based on sparse representations to recover quantitative and remote sensing products that are contaminated by thick clouds and consequent and attendant shadows. Despite its great success to achieve remarkable performance in the field of human action recognition, the aforementioned approaches have several limitations. In fact, they have difficulty in handling lengthy videos containing high levels of illuminations and temporal occlusions, and their capacity was stymied to effectively model and learn long-term temporal information. Furthermore, the requirement of an engineering process to extract features, get representations and build vocabulary is labor-extensive (Table 1).

2.2 Deep learning-based approaches

Owing to the cited limitations of handcraft-based approaches, many researchers focused their works on deep learning-based approaches. These latter have shown significant improvements in several domains such as object tracking [23,24], video saliency [25,26], image cropping [27], mental activity observation [28] and image reconstruction [29].

For visual object tracking, Dong et al. [23] proposed a quadruplet Siamese deep network that uses the potential connections among the training instances to achieve more powerful and robust representations for one shot learning. The authors used a combination between triplet and pair loss by automatically adjusting weights to improve the perfor-

Table 1 Related works on handcraft-based approaches in HAR

References	Year	Feature extraction	Feature classification	Dataset	Acc (%)
Asadi et al. [15]	2018	Large margin nearest neighbor (LMNN) + efficient match kernel (EMK)	Linear SVM	multicolumn1lMSR action 3D	95.6
				MSRDailyActivity3D	85.0
				MSR gesture 3D	97.0
				3D action pairs	100
				UT kinect	97.0
Oreifej et al. [14]	2013	A histogram of oriented 4D surface normals (HON4D)	SVM	MSR action 3D	88.9
				MSR gesture 3D	92.4
				3D Action Pairs	96.6
Hu et al. [19]	2009	MHI and HOG	SMILE-SVM	CMU action	–
				Système de surveillance dans les centres commerciaux	–
Klaser et al. [13]	2008	Histograms of oriented 3D spatiotemporal gradients	SVM	KTH	91.4
				Weizmann	84.3
				Hollywood	24.7
Willems et al. [18]	2008	SURF3D	SVM	KTH	84.2
Bay et al. [17]	2007	SURF	Bayes classifier	Caltech background and airplanes set	–
Scovanner et al. [12]	2007	3D SIFT descriptor	SVM	Weizmann	82.6

mance. To get around the drift problems caused by partial occlusion and appearance deformation, Liang et al. [24] proposed a local semantic Siamese network to learn local semantic features during the offline training for fast object tracking. In fact, they added a classification branch and a residual channel attention block into the classical Siamese framework. To further enhance the representation of features, a focal logistic loss is designed to mine the hard-negative samples. During the online tracking, the classification branch is removed and an efficient template updating strategy is applied to handle long-term object deformation. For fast and efficient video saliency detection, Wang et al. [25] proposed a deep learning model by using convolutional neural networks. The proposed deep video saliency network involves two parts, namely static saliency network and dynamic saliency network, which are considered to capture static and dynamic saliency information. The saliency estimated from the static network is fused into the dynamic network to produce accurate spatiotemporal saliency result. Lai et al. [26] proposed a spatiotemporal residual attentive neural network to predict dynamic attention from limited data. The proposed network emphasizes two parallel DNN streams to capture and predict spatial and temporal saliency features. To learn the comprehensive spatiotemporal saliency representation, the saliency features of two network streams are fused by incorporating dense residual crossconnections among different layers. To further enhance the spatiotemporal saliency representations, the authors integrate a composite attention module to learn the local and global attention priors. To model the temporal attention transitions across video frames from limited data, a lightweight recurrent network convGRU is introduced to the network. Deep learning is also used on photograph cropping. This technique is used to further polarize the image focus toward the salient region which can be defined as the most visibly bringing out region of an image. Studying this problem, [27] proposed a deep approach designing a model composed of two subnetworks: an attention box prediction (ABP) network and an esthetic assessment (AA) network; both of them is designed to share the same multiple convolutional feature maps of initial convolution layers. The ABP network derives initial cropping as an attention bounding box. From a few cropping candidates generated around the initial cropping,

the AA network selects the final cropping window with the best esthetic quality among the candidates. The authors leveraged attention prediction and esthetic assessment to produces high-quality cropping results.

These implemented deep learning models have also been investigated for action recognition applications. Deep learning-based approaches [11,30–33] are end to end trainable, and they can directly be applied on the raw data frame. To deal with the challenge of action recognition, [30] implemented 3D convolutional networks in time axis to extract spatial and temporal features. Karpathy et al. [11] modified the CNN architecture to a multi-resolution framework separated in two streams: The context stream learns low-resolution features, and the fovea stream learns high-resolution features by applying center crops regions in the middle portion of the image. In each video sequence, CNNs was applied to multiple frames to obtain temporal information using three types of fusion (early, late and slow); nevertheless, this approach does not provide considerable improvement when compared to single-frame models. Simonyan et al. [31] proposed a two stream CNN architecture that incorporates two feature types: The first one is spatial stream that takes RGB frames as input data, and the second is a temporal stream that takes as input dense optical flows. Since features extracted from optical flow images contain only short-term temporal information, adding it to a framework does not enable to learn long-term temporal dependencies between frames. Tran et al. [32] propose in their research work, a deep network named 3D convolutional framework (C3D) that enables to extract temporal features in an end-to-end structure which evaded the requirement of pre-computing optical flow features. Nevertheless, C3D approach covers only a short interval of the video sequence. To represent the temporal dynamics between frames of the entire video, recurrent neural networks (RNNs) have been employed in the task of human activity recognition (HAR)-based video. RNNs are powerful and robust type of neural networks that are used to find and extract hidden patterns of data in both space and temporal domains. In RNN, the data are processed sequentially and go through a loop so that when the system makes a decision, it takes into account the current input and the learned inputs received previously. The majority of the state-of-the-art approaches [34–40] that have combined CNNs and RNNs networks for human activity recognition have obtained impressive results. Nevertheless, because of the huge number of computation's parameters and the disappearance of preliminary inputs after few layers, the problem of vanishing gradient has appeared. An effective method have been proposed as a solution to this problem, which is LSTM (long short-term memory) [35,37,41] that are able to acquire long-term dependencies and integrate multiplicative logic gates allowing to store and access relevant information over long intervals, thus reducing the impact of the vanishing gradient problem. Combining CNN and LSTM to extract spatial and temporal characteristics is a technique that has attracted attention to deal with computer vision tasks. For instance, an implementation of daily living activities (DLA) recognition using deep networks is developed in [42] proposing two deep learning approaches that exploit LSTM to learn long-term temporal dependencies. The first approach is a multi-scale LSTM (MT-LSTM) model which combines three LSTMs to detect temporal dependencies of the activity from preprocessed features of skeletal data. The second is a CNN-LSTM model which combines convolutional and recurrent networks to extract spatial and temporal information from raw data. Reference [37] proposed and evaluates two types of deep neural network architectures to merge spatial and temporal image information over longer periods of time. The first one uses a CNN network to explore high-level features of frames and improves the classification accuracy by increasing the number of frames. As CNN-based approaches are able to extract only visual appearance features, and lack the capability to model a long-term temporal information, a second model was proposed to explicitly represent the video as a sequence of frames and connect the output of the CNN network to the recurrent LSTM architecture. Therefore, for better action recognition, reference [43] suggested CNN and bidirectional LSTM networks to reduce the complexity and redundancy. The CNN model extracts features from the sixth frame of videos, and the BD-LSTM network is used to learn the long-term sequential information from features of the lengthy videos. The BD-LSTM model is composed of two layers each having forward and backward passes. Reference [44] used two stream ConvNet to extract spatial and temporal features using ResNet-101; this work concatenates spatial and temporal features to construct feature matrices which are used as data input to a temporal segment LSTM or a temporal inception for activity prediction to better exploit temporal information. Reference [45] proposed a system that combines bidirectional gated recurrent unit based on recurrent neural network with a 3D CNN in a voting manner. The resulting RNN and CNN features are then fused and fed into an SVM classifier for action prediction. In [46], Lieyun Ding et al. implemented a deep learning hybrid model which combines CNN and LSTM neural networks to recognize workers' unsafe actions. In fact, CNN generates spatial features from each video frame, which are used as input data to the LSTM model to learn about their temporal dynamics. The authors in [47] proposed a two-stream approach to recognize human action from multimodal video data involving RGB images and articulated poses; a convolutional model takes 3D tensor data as input to process the pose stream and a spatiotemporal attention mechanism is used to manage the RGB flow. Reference [48] incorporated a deep three-dimensional convolutional network (C3D) and LSTM networks to capture spatiotemporal dynamics over

long range. The C3D network models motion and appearance information integrating temporal information, and then, the LSTM is used to encode features and to classify the activities. Reference [49] proposed a deep architecture that employs LSTM network to capture long-term temporal information that evolves with pose change and CNN to focus on static appearance information. Scores of LSTM-based skeleton and CNN-based appearance classifiers are fused to obtain the final score for activity classification (Table 2).

In the light of these earlier works, the present paper proposes a simple and effective method to handle with the human action recognition issue. The developed framework aims to incorporate the residual convolutional neural network with the recurrent neural network to acquire the temporal information from time series data. In this paper, we used a residual-based CNN network to generate global contextual feature representation, enabling the use of existing CNN models namely ResNet-101 directly applied on video data using fine-tuning techniques. Our model provides effective feature representation for spatial and temporal sequential data by combining CNN and RNN networks; in fact, RNN uses the extracted residual CNN features as input and provides stronger encoded features that take into account long-term temporal dependencies of the whole video sequence input data. The resulting spatiotemporal feature vector is inserted into the softmax layer that operates as classifier by generating probabilities which will be used to classify actions and then to recognize them by predicting the corresponding label of the video sequence. Experimentally, the proposed approach shows competitive results compared to state-of-the-art methods when applied on two publicly available daily activity benchmarks.

3 Methodology

In this section, we describe our deep feature representation method for HAR in detail. We first introduce the deep residual neural networks and how they are exploited for residual feature extraction. Secondly, we introduce LSTM networks and the technique to extract temporal features taking into account temporal information dependencies between video frames.

3.1 Data augmentation techniques

Following [11,16], we benefit from the data augmentation method to expand our datasets artificially by applying some transformations and geometric deformations to original images allowing to produce transformed images. This prevents the model from overfitting by increasing the number of frames having the same legends and help to better discriminate between pertinent characteristics. Indeed, before introducing data to the network, we apply preprocessing operations to all frames. The first kind of data augmentation consists of making image reflections on the left direction, so that pixel values were reflected along the boundaries of the frame. The second stage of preprocessing consists in translating the images by moving them on the horizontal and vertical axes.

3.2 Temporal residual representation

3.2.1 Deep residual neural networks

Recently, deep residual networks have shown persuasive performance and acceptable convergence behaviors on several experiments carried out on large scale image recognition challenge, such as ImageNet [50]. These studies claims that residual frameworks are very helpful to attenuate the degradation problem when training deeper networks, since they involves more than hundred of layers by announcing a novel architecture which includes shortcut connections making the residual networks more easy to optimize than the plain networks. The difficulty in learning such deep networks is particularly related to the backpropagation of the gradient. Furthermore, with deeper networks, the gradient is lower for updating the weights of the lowest level layers (first layers), so too deep architecture does not really update these layers. The main idea proposed in ResNet [7] is to utilize residual connections enabling beneficial optimization for very deep neural networks. A residual connection materializes the task to pass the input in two convolution filters and also to pass the same input directly to the following layers. And this is done by calculating the sum of the results of the two convolutional layer and the input value. By using this structure, the authors demonstrate the interest of learning very deep neural networks by their performances and facilitate their efficient training.

Before the development of the residual network architecture, it was impossible to train a network with more than 25 layers. As the layers became deeper and the gradients became smaller, the performances were inevitably degraded: The error was no longer propagated correctly, and the update of the weights was directly affected. Residual neural networks have made it possible to go beyond this limitation. Their architecture allows the creation of very deep neural networks, with better accuracy than those with linear architectures because they have the ability to extract more information and thus have a more advanced analysis of images.

Residual blocks A residual block is considered as the basic unit in a residual network, and each residual block contains a residual branch and an identity mapping. The corresponding formula is represented as,

Table 2 Related works on deep learning-based approaches in HAR

References	Year	Feature extraction	Feature classification	Dataset		Acc (%)
Ma et al. [44]	2019	CNN (5 conv layers)	Temporal segment LSTM or temporal ConvNet	UCF101		95.7
				HMDB51		69.0
Arif et al. [48]	2019	3D ConvNet (C3D)	LSTM encoder–decoder	UCF101		94.0
				HMDB51		70.1
Kim et al. [39]	2018	CNN-GRU	Softmax	Nine gesture based MSR dataset of Microsoft		97.0
Baradel et al. [47]	2018	CNN + LSTM	Stream fusion by softmax classifier	NTURGB + D		87.7
				SBU kinect interaction		94.1
				MSRDailyActivity3D		90.0
Ding et al. [46]	2018	CNN (5 conv layers) + two LSTM layer	Softmax classifier	200 pictures	With two types of labels	97.0
					With four types of labels	92.0
Ullah et al. [43]	2017	CNN (5 conv layers)	BD-LSTM (2 LSTM layers)	You Tube		92.8
				HMDB51		87.6
				UCF101		91.2
Veeriah et al. [36]	2015	HOG3D	LSTM (dRNN)	MSRAction 3D		94.0
				KTH1		93.9
				KTH2		92.1
Wu et al. [38]	2015	CNN + LSTM	Regularized fusion network	UCF-101		91.3
				CCV		83.5
Yue-Hei Ng et al. [37]	2015	LSTM (5 stacked layers)	Softmax	Sport-1M	Hit@1	–
					Hit@3	73.1
					Hit@5	90.5
			Softmax	UCF-101	With optical flow	88.0
					Without optical flow	82.6
Simonyan et al. [31]	2014	CNN (M-2048 architecture)	Softmax classifier	UCF-101		49.0
			SVM classifier	HMDB-51		46.6
				HMDB-51 and UCF-101		55.4
Karpathy et al. [11]	2014	Multi-resolution CNN (5 conv layers)	Multilayer NN with ReLU + softmax	UCF-101		63.3

$$x_{l+1} = x_l + F(x_l, w_l) \quad (1)$$

where x_l and x_{l+1} match, respectively, to the input and the output of the lth residual block, F represents a residual function and w_l is a set of weights, respectively, associated with the block. A residual network is composed of residual blocks stacked sequentially.

3.2.2 The ResNet-101 model

ResNet-101 is a model that is easy to implement and particularly well suited for different types of recognition problems. It has been pre-trained on the ImageNet database, which contains more than 14 million images classified in more than 20,000 category and available by the image analysis community of Stanford Vision Lab. This model is composed in total of 347 layers corresponding to 101 layer residual network, has the particularity of introducing residual connections and can classify images into 1000 object categories. The ResNet-101 architecture is composed of five convolutional layers, two pooling layers and two fully connected layers. Moreover, the Residual Network ResNet-101 model based on the work of [7] is required because the network architecture requires a very large number of parameters. Learning all these parameters using a small database is a significant challenge that causes an enormously waste of time. Unlike convolutional neural networks that have a linear architecture (a stack of layers for which each output is only connected to the next layer) (see architecture A of Fig. 1), in a residual network, the output of the previous layers is connected to the output of the new layers to pass them both to the next layer. A schema is required (see architecture B of Fig. 1): In ResNet model, every layer is composed of several blocks. ResNets go deeper by increasing the number of operations within a block; however, the total number of layers remains unchanged. An operation consists of a convolution, a batch normalization and a rectified linear unit (ReLU) activation function, apart from the operation of the last block that does not apply the ReLU activation. At the heart of this model, the main idea is that the identity function must be added at every additional layer. This denotes that when we train a new additional layer into an identity mapping: $f(x) = x$, we obtain an effective model as the initial one. The new model is able to give better solution to adapt the training dataset, and so, to extract deeper sparse and pertinent residual representation of spatial features from the action video frames, the additional layer can facilitate reducing the training errors.

3.2.3 Deep residual feature extraction

In this paper, the residual network ResNet-101 is used with pre-trained parameters from ImageNet database and applied to extract sparse and pertinent residual representations of spatial features from video frames of each video sequence. The architecture is composed of several ResNet blocks with three layer deep composed of five composite convolutional layers including small kernels sizing by 7×7, 1×1 and 3×3, The output is obtained from the average pooling operation applied to the final feature map of the network followed by the fully connected layer "fc." In fact, the 2048-dimensional features resulting from the last average pooling layer "pool5" with $1 \times 1 \times 2048$ activations are used as input to the fully connected layer "fc" with $1 \times 1 \times \text{Nbre}_{\text{InputVideo}}$ activations. The finally yielded $\text{Nbre}_{\text{InputVideo}} \times 2048$ is considered as residual features generated from the reused pre-trained model in a feed forward pass. In fact, each sequence video was considered as a class of activity apart. So, we have as many classes as number of video sequences in the database. Each video was separated into many clips. Video frames are picked to be processed randomly, as they were represented temporally during the course of the action. For each input frame, we calculate the output of neurons connected to the local regions in the residual layers. In this input volume, we calculate the product between weights and the small region to which neurons are connected and an identity function must be added at every additional layer like one of its elements. This denotes that when we train the new additional layer into an identity mapping: $f(x) = x$, we obtain an effective model as the initial one. The new model is able to give better solution to adapt the training dataset, and so, to extract deeper sparse and pertinent residual representations of spatial features from the action video frames, the additional layer can facilitate reducing training errors. A feature vector of dimension Nbrinput_video $\times$ 2048 is the outcome of the fully connected layer which represents the deep residual feature of each sequence video from the database. Figure 2 summarizes the proposed method for the residual features extraction.

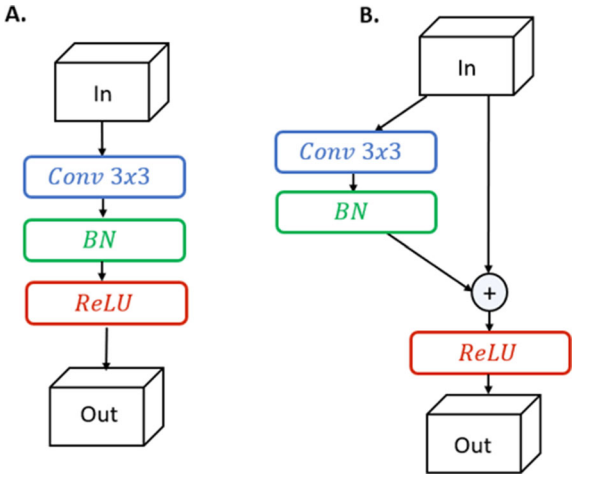


Fig. 1 Extract from the architecture of a convolutional neural network without **a** and with **b** residual connection

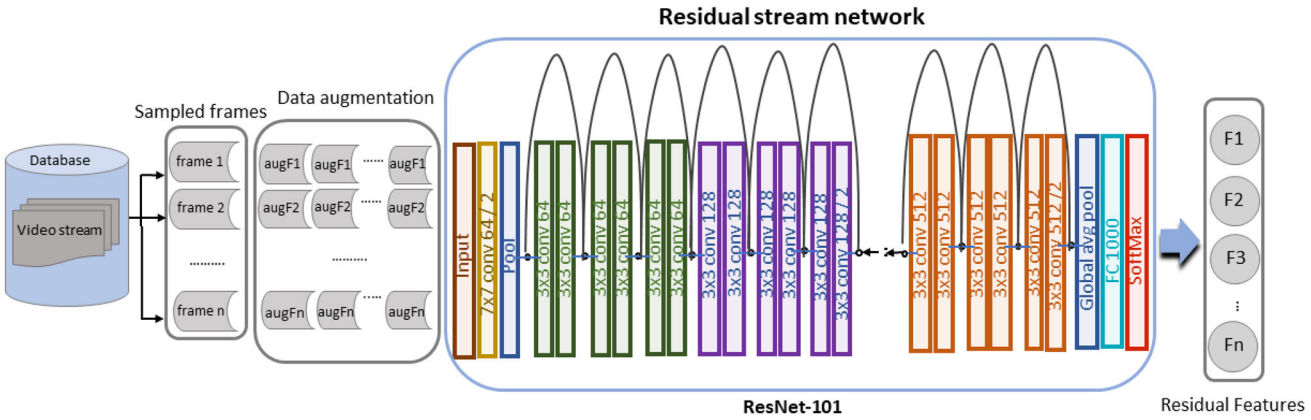


Fig. 2 Proposed residual feature extraction methodology

3.3 Deep temporal feature encoding

3.3.1 Recurrent neural network

To examine the hidden patterns and encode spatial extracted features temporally, recurrent neural network (RNN) is required. A recurrent neural network models the temporal dynamics by generating the current hidden state h_t and the output y_t utilizing its current input x_t and the previous hidden state h_{t-1} via the following equations:

$$h_t = F(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2}$$

$$y_t = F(W_{hy}h_t + b_y) \tag{3}$$

where W_{xh}, W_{hh} and W_{hy} stand, respectively, for the input-to-hidden, the recurrent hidden-to-hidden and the hidden-to-output connections. b_y and b_h are biases terms for output and hidden states, respectively. F is an element-wise nonlinearity, such as a sigmoid, hyperbolic tangent or a rectified linear unit. To handle video sequences and understand an action context, we must explore the visual information restricted to each sequential video frame. RNN is able to deal with such problems, but it can be difficult to train long-range sequences, because it frequently forgets the earlier sequences information's, which are due to the vanishing or exploding gradients problems [51] that can be produced when propagating gradients down over many RNN layers. The vanishing gradient occurs when it tends to zero due to n small derivatives that are multiplied together across T indices of time. The exploding gradient arises over exponential increase by multiplying gradients repeatedly through all network layers. This phenomenon will eventually lead to a totally unstable network. LSTM has been considered to overcome the vanishing and exploding gradient problems by integrating memory units to learn long-term temporal dynamics over sequential frames.

3.3.2 Long short-term memory (LSTM) model

LSTM is a variety of recurrent neural networks that contain memory cells that facilitate to learn long-term dynamics, and conserve data information through time. Furthermore, LSTM network announces an exceptional structure that uses three gates (input, forget and output) to supervise and update the cell memory's state and to manage long-term temporal dependencies among consecutive frames. These gates make adjustments by a sigmoid unit to track the information flow during the training phase. We adopt LSTM network for human action recognition because it is beneficial to model sequential data in computer vision challenges. First, it allows the end-to-end fine-tuning in straightforward structure. Second, when manipulated with sequential data such as video sequences, LSTM is not closed to a fixed-length data of the network architecture. Figure 3 (right) demonstrates a basic structure of LSTM unit. An LSTM neuron is able to select at each time step, the type of operation (read, write or reset) that will be applied to the memory cell through the mechanism of gates, and the latter are used to control the information received by the cell. This technique helps LSTM to retrieve and retain the information over many time steps. An LSTM unit accommodates an input gate i_t, an output gate o_t, a memory cell c_t and a forget gate f_t. For each time step t, LSTM performs the memory cell update by the following equations:

$$i_t = \delta(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{4}$$

$$f_t = \delta(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{5}$$

$$o_t = \delta(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{6}$$

$$g_t = \delta(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{7}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \tag{8}$$

$$z_t = h_t = o_t \otimes \tanh(c_t) \tag{9}$$

x_t, h_t, c_t and z_t refers to the input vector, the hidden state, the cell state and the output at the tth state, severally. The

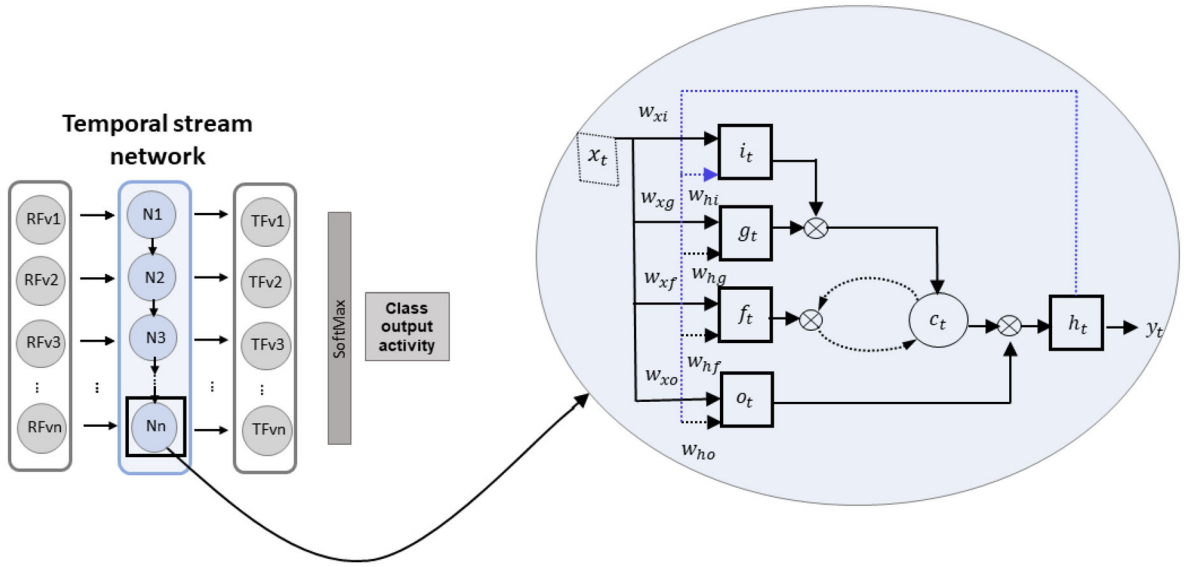


Fig. 3 Architecture of the temporal stream network (left). The LSTM layer is consisting of LSTM cells. Each LSTM cell is structured as (right)

output z_t is dependent on the hidden state h_t, whereas h_t is decided by the cell state c_t, and its prior state h_{t-1}. W and b are the LSTM network parameters' corresponding to the input vector' weights and the bias, respectively. $\delta(x)$ denotes a sigmoid function defined as $\delta(x) = (1 + e^{-x})^{-1}$, tan$h$ is known as an activation function and $\otimes$ is specifying the element wise for the multiplication operation. The cell memory state c_{t-1} and the output z_t are assessed iteratively to extract the long-range dependencies between sequential frames. x_t is the new representation of the spatial feature vector after transformation, which is employed as input to the LSTM network. The forget gate f_t rubs out less significant information from the cell memory c_{t-1}, and the output gate o_t determines the amount of information from the cell memory c_t to be transmitted to the hidden state h_t. g_t is a function of the actual input frame and the precedent hidden sate h_{t-1}. The hidden state is calculated using the memory cell c_t and the activation function tan$h$.

3.3.3 Temporal feature extraction

Our DTR-HAR model works by processing each input frame Fr_t on a feature transformation φ parametrized by ϱ to create a fixed-length vector representation φ_t as shown in this equation:

$$\varphi_t = \varphi_\varrho(Fr_t) \tag{10}$$

This vector will be considered as input for the LSTM network. Furthermore, each visual input frame must be encoded function as the residual CNN output features to generate new encoded feature representation that will be fed into LSTM to extract temporal features needed for the activity prediction. This step consists to transform an input data X_t to a corresponding activation h for the last specific fully connected layer "fc" of the trained convolutional residual network explained in details in Sect. 3.3.1. Since the obtained matrices are multi-dimensional, we flattened them into one-dimensional vector by concatenating all matrix weights to obtain a new feature vector that will be fed to the LSTM which collects the temporal dependencies information's between all frames of the sequence video. Differently from the previously used CNN model, the actually considered LSTM model is composed of a single layer where the number of neurons corresponds to the number of features extracted by CNN and transformed by the φ function as presented the equation 10. Generally, the LSTM is parametrized by the weights and biases corresponding to W and b of the input and the hidden layers generating an output y_t of the input x_t and the anterior time step hidden state h_{t-1} in a way that each output is inferred by the previous helping to update the current hidden state h_t at each time step. Finally, the output of the LSTM was considered as input to a fully connected layer with a softmax activation function to predict the activity performed in the sequence video. This softmax layer is employed to attain the class scores for the given video. The prediction distribution $P(Y_t)$ at the current time step t is achieved by calculating the average of the probabilities scores inferring the class activity, following Eq. 11:

$$P(Y_t^{\vartheta} = 1) = \text{softmax}(Y_t) = \text{softmax}(W y_t + b_t) \tag{11}$$

where W and b_t stand for the trained parameters designing weights and biases of the LSTM model and $\vartheta \in \vee$ is designing the prediction. The developed model is considered

advantageous by its flexibility and its independence from the length of the video sequence to be processed.

4 Implementation details

Base architecture Our spatial residual ConvNet stream is based on ResNet101 architecture. More specifically, our model takes clips with a size of $224 \times 224 \times 3$, and 224 is the height and width of input frame. We chose such input resolution to facilitate the reuse of existing pre-trained image classification network models without requiring to retrain the network from scratch. This prevents the relatively used small datasets from overfitting. The spatial stream (residual CNN) is pre-trained on a classification training subset of the ImageNet, i.e., the large visual database is designed for use in visual recognition issues [50]. The resulting 2048-dimensional features generated from the average pooling layer "pool5" with $1 \times 1 \times 2048$ activation are used as input to the fully connected layer "fc" with $1 \times 1 \times \text{Nbre}_{\text{InputVideo}}$ activation. A spatial feature vector of size $\text{Nbre}_{\text{InputVideo}} \times 2048$ is resulting from the last fully connected layer.

For the temporal stream, each video sequence has been split into frames which are in turn divided into train and test sets. Activations are applied to each video sequence subset using the last fully connected layer of the spatial residual convNet of size $\text{Nbre}_{\text{InputVideo}} \times 2048$. New representations of each video sequence frames are collected into a single vector which is considered as input for LSTM model to extract temporal features by sequencing them taking into account the temporal dependencies of activities in the frame sequences and to finally output the activity class of each video. Since the used action recognition datasets are considerably small for the training which increases the risk of overfitting, data augmentation operators detailed in Sect. 3.1 are applied in the training stage to improve the performance of our network architecture in such a way that the training dataset was artificially augmented at each iteration. Each sequence video from the datasets needs to be split, and the obtained video frames are required to be processed. Therefore, N individual frames are inputted into the residual ConvNet network which are then connected to temporal network composed of one single-layer LSTM with the number of hidden units corresponding to the number of features reproduced by the residual ConvNet model.

Hyper-parameter optimization Our spatial residual ConvNet model is strongly initialized with ImageNet pre-trained weights for faster training. The spatial network is trained end to end by optimizing the crossentropy cost function using "stochastic gradient descent" (SGD) and backpropagation. We used a learning rate of 0.0001 and a mini batch size of 50, and the network is trained for 6 epochs for two used datasets. Our residual ConvNet model is trained for more than 16000 iterations for three days. The network parameters converged after around 1000 iterations with one epoch. For the temporal stream, we employed "Adam" parameter update algorithm for optimization. For MSRDailyActivity3D dataset, the learning rate factor is set to 0.001 to fine-tune parameters' network along 100 epochs and the batch size were set to 32. For CAD-60, the network was trained for 300 epochs with learning rate factor set to 0.001 and a mini batch size of 16. For both two networks, the momentum is set to 0.9. The temporal model is trained for more than 1000 iterations for one hour. Our DTR-HAR model was conducted on a 2.6 GHz machine (Intel Core i7-6700HQ CPU) with 8GB of DDR4 RAM using MATLAB 2018a. We trained and tested our model on a NVIDIA GeForce GTX 960M with 4GB GDDR5 GPU machine.

5 Experimental results

This section describes the experimental setup of the training process and the experimental results of our proposed method. We have carried out several experiments in order to be able to make a good comparison of the performances of the method proposed in this work and those of the two state-of-the-art datasets are evaluated using the proposed deep neural network-based model for activity recognition, namely CAD-60 [52] and MSRDailyActivity3D [53] datasets.

5.1 Dataset description

Cornell activity dataset CAD-60 [52] is an activities dataset recorded by a Microsoft Kinect sensor. It contains 60 RGB-D videos with 12 different actions performed by 4 subjects. The performed actions are: "rinsing mouth," "brushing teeth," "wearing contact lens," "talking on the phone," "drinking water," "opening pill container," "cooking (chopping)," "cooking (stirring)," "talking on couch," "relaxing on couch," "writing on whiteboard" and "working on computer." Each video is described with RGB, depth and skeletons modalities.

MSRDailyActivity3D dataset [53] This dataset was captured by a Kinect sensor, to deliver depth maps, RGB video and skeletons modalities. In total, it contains 320 RGB-D video samples of 16 daily activities executed by 10 different subjects. All activities are performed in the living room, those are: "Drink," "Eat," "Call cellphone," "Read book," "Use laptop," "Use vacuum cleaner," "Write on a paper," "Sit still," "Toss paper," "Play game," "Cheer up," "Lay down on sofa," "Play guitar," "walking," "Play guitar," "Stand up" and "Sit down." Each activity is executed twice by the same person: one in the standing and the other in the sitting position.

5.2 Comparisons analysis

For analysis purpose, many evaluation parameters are calculated such as accuracy, precision, recall and F-score are calculated using the following mathematical equations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100 \tag{15}$$

In general, positive means identified and negative means rejected. So, in the above-stated equations, TP (true positive) corresponds to instances correctly identified, FP (false positive) for instances incorrectly identified, TN (true negative) for instances correctly rejected and FN (false negative) for incorrectly rejected instances. To evaluate the performance of our proposed DTR-HAR approach, we compare it with different existing state-of-the-art human activity recognition methods. The activity recognition performance of the proposed DTR-HAR approach and the competing methods on CAD-60 and MSRDailyActivity3D datasets are summarized in Table 3, in which the overall accuracy' results of all the competing methods are selected from the corresponding publications. Table 3 shows that the proposed DTR-HAR approach achieves competitive results compared to all recent state-of-the-art methods on both CAD-60 and MSRDailyActivity3D datasets, thus justifying the effectiveness and the good generalization of the proposed model. Datasets were considered to involve daily living activities performed by different persons. We have split each dataset to training and test subsets to evaluate the performance of our approach and, so, the quality of our model by calculating the corresponding accuracy value. Each video was sampled into frames which serves as input to the residual ConvNet network. The pre-trained ResNet-101 was used to learn spatial features and to fine-tune hyper-parameters.

5.2.1 CAD-60 dataset results

In Table 3, our deep temporal residual (DTR-HAR) model is compared to different state-of-the-art methods. The 5-CNN streams approach [54] is lower than our proposal for 1.18 points, although it is a multimodal approach involving RGB, depth and skeleton as input modalities, whereas our proposed model use only RGB frames as input modality. This confirms that our proposed model works considerably well on CAD-60 dataset.

The confusion matrix represented in Fig. 4 entails the accuracy of each action class; i.e., each row corresponds to the predicted class and each column designs the actual class. More than half of the activities, (especially, cook (chopping), Open container, Open pill container, Relax, Rinsing, still, talking couch, talk phone, wear lens, wear cont lenses and work computer) were 100% correctly classified and the other are classified with a high level of confidence, which proves the efficacy and the robustness of our proposal.

Most of the activities are correctly classified with a high confidence level. As illustrated in Table 3, our proposed model performed well and achieved superior accuracy with similar state-of-the-art approaches, which is highlighted in bold text.

The action class brush teeth is misclassified by 25% as shown in Fig. 4. It was sometimes recognized as Rinsing action class. This misclassification is explained by the similarity of both fine-grained actions with subtle motion variations tacking place in the same background, which are describing a person carrying an object in his hand, in the direction of his mouth. Figure 5a shows a person who is brushing his teeth, and Fig. 5b shows a person who is rinsing his mouth. The object that was not clearly distinguished in both frames is the toothbrush in the brush teeth action class and the glass in the rinsing action class. In addition, both activities have the same hand movement, thus inducing confusion between the two classes of action. A similar comment applies to the action class Drink which is misclassified by 25% and is confused with the action class talk phone. This misclassification is due to the similarity of the two actions which involve a person taking an object in his hand and the position of the hand is raised toward his head. For the action class Random, it was misclassified by 25% and is confused with Write board. This can be explained by the presence of a white board behind the person performing the action, frames involving a person very close to the white board are then classified as White board action. Figure 6a shows a person who is performing a Random action, and Fig. 6b illustrates the case of a person who is writing on a white board. Frames of the two actions which involve a person in front of a white board are making the confusion, because of their repetition in both actions within Random and white board. In addition, the cook stirring action class is confused with -work computer-, since in both action classes the person doing the action is in front of a table and he is manipulating an object using his hands.

We evaluate the performance of our system by exploiting various evaluation metrics for each activity separately. The obtained results are presented in Fig. 7 which is in fact a histogram illustrating the precision (in blue), Recall (in orange), F1-score (in gray) and accuracy (in yellow) for each activity of CAD-60 dataset. The majority of activities have been correctly classified (with precision, recall, F1-score and

Table 3 Results on CAD-60 and MSRDailyActivity3D datasets using RGB videos

Dataset	Method	Accuracy (%)
CAD-60	STIP [55]	62.50
	MEMM-HOG* [52]	64.20
	Multi-level depth fusion* [56]	67.40
	Actionlet ensemble^{+} [53]	74.70
	Object affordance^{+} [57]	71.40
	MSLF^{+} [58]	80.36
	JOULE-SVM^{+} [59]	84.10
	4-stream CNN [60]	89.05
	5-CNN streams [54]	90.00
	DTR-HAR (ours)	**91.18**
MSRDailyActivity3D	CNN-LSTM^{+} [61]	63.10
	RGB + CS-Mltp + SVM [62]	65.63
	Pose-driven attention [63]	76.64
	IPM [64]	83.30
	DSCF^{+} [65]	83.60
	P-CNN + kinect + pose machine*$^{+}$ [66]	84.37
	RGGP + fusion^{+} [67]	85.60
	Actionlet ensemble^{+} [53]	85.80
	MSLF* [58]	85.95
	BHIM [68]	86.88
	IPM + joints [64]	89.30
	DTR-HAR (ours)	**91.56**

Bold indicates the results of our method for each evaluation metric of performance
*Corresponds to methods requiring depth modality. $^{+}$Corresponds to methods requiring skeleton modality

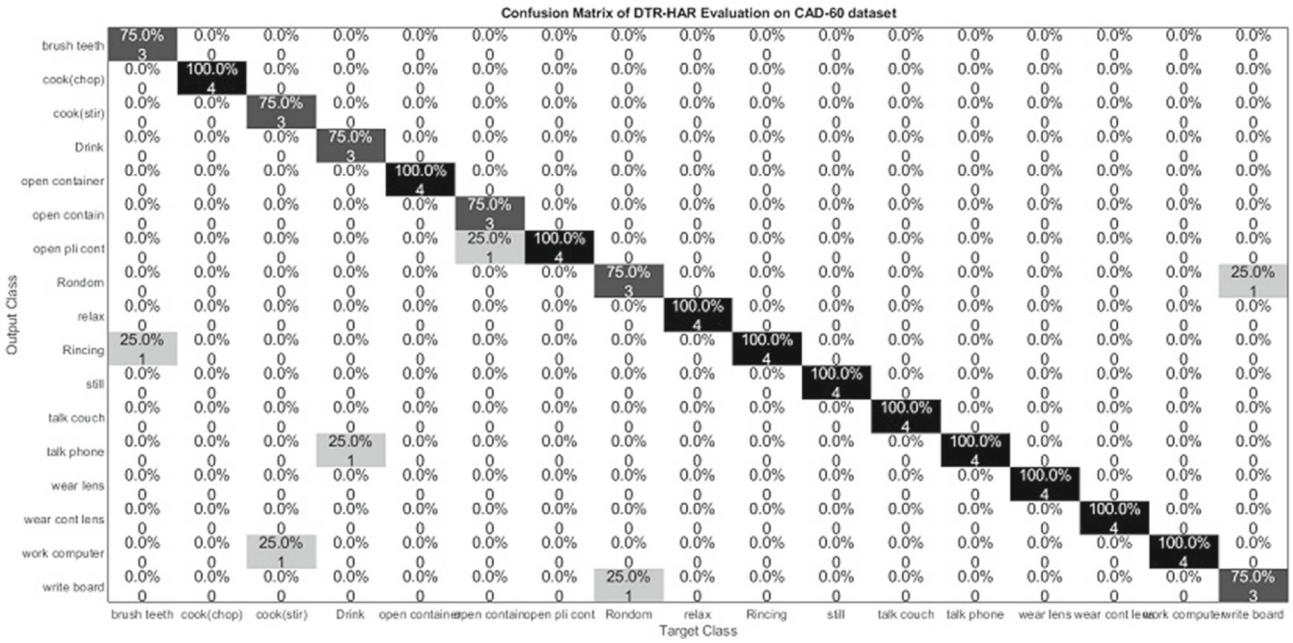


Fig. 4 Confusion matrix of DTR-HAR evaluation on CAD-60 dataset

Fig. 5 **a** A brushing teeth person from frames set of class brush teeth, **b** a rinsing mouth person from frames set of class Rinsing

Fig. 6 **a** A random person from action class Random, **b** a writing board person from action class white board

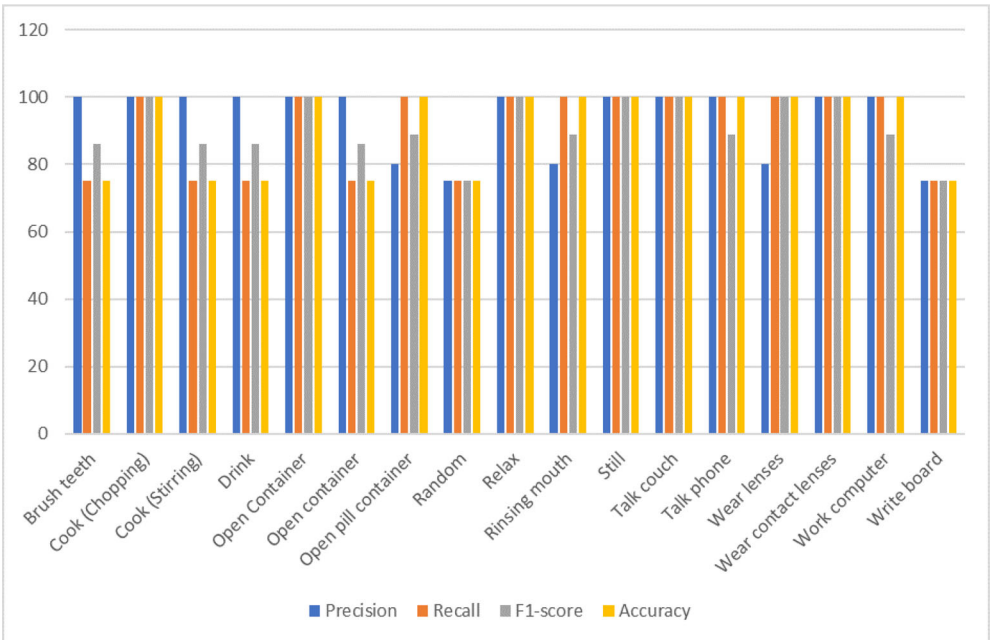


Fig. 7 Performance evaluation metrics by activity in CAD-60 dataset

Table 4 Precision and recall of the DTR-HAR model on CAD-60 dataset

Activity	Precision (%)	Recall (%)	F1-score (%)
Brush teeth	100.00	75.00	86.00
Cook (chopping)	100.00	100.00	100.00
Cook (stirring)	100.00	75.00	86.00
Drink	100.00	75.00	86.00
Open container	100.00	100.00	100.00
Open container	100.00	75.00	86.00
Open pill container	80.00	100.00	89.00
Random	75.00	75.00	75.00
Relax	100.00	100.00	100.00
Rinsing mouth	80.00	100.00	89.00
Still	100.00	100.00	100.00
Talk couch	100.00	100.00	100.00
Talk phone	100.00	100.00	89.00
Wear lenses	80.00	100.00	100.00
Wear contact lenses	100.00	100.00	100.00
Work computer	100.00	100.00	89.00
Write board	75.00	75.00	75.00
Overall average	**92.00**	**91.00**	**91.00**

Bold indicates the results of our method for each evaluation metric of performance

accuracy are at 100%). Only the writing board and Random are the worst in their classifications, with evaluation metrics not lower than 75% (which are still acceptable). This misclassification is explained by their great similarity in many details when carrying out the action.

Precision, Recall and F1-score are calculated and summarized in Table 4, proving the good results achieved by the DTR-HAR model on CAD-60 dataset, detailing the performance attained for each activity class. This can be explained by the good ability to discriminate between the four persons realizing the actions and confirms that our DTR-HAR model is able to treat temporal patterns in order to predict actions.

5.2.2 MSRDailyActivity3D dataset results

Table 3 displays the performance comparison of our DTR-HAR framework with state-of-the-art approaches on MSRDailyActivity3D dataset. Although the DTR-HAR framework involves only the RGB modality type, we can guarantee that our model outperforms the other state-of-the-art approaches whether or not combining multiple modalities. This demonstrates that the learned representations can generalize across domains. The work of [61] has obtained bad results in this dataset although it was based on the combination of two deep neural networks that was CNN combined with LSTM; however, the used CNN is not based on transfer learning concept. Learning the model and fine-tune hyperparameters on a large-scale dataset is very efficient to obtain high performance level on a small one.

Based on the confusion matrix results in Fig. 8, the majority of the activities were correctly predicted since they were highly discriminated. Nevertheless, in some cases, some actions were confused with each other due to the similarity of fine-grained actions with subtle variations in the same background. Indeed, a person can exist in similar positions during the realization of two different activities, such as with drink and call cellphone action classes. The action class drink is confused with call cellphone by 5.0%. This confusion is explained by the fact that the person performing the two actions holds an object in the hand by bringing it closer to his head to perform the desired action, once the action is performed, he lowers his hand. Here, the object handled for the two fine-grained actions was not correctly detected for the two similar actions drink and call cellphone.

Figure 9 displays a person carrying a glass in his hand to drink, when finished, he lowers his hand down and Fig. 10 illustrates the case of a person holding a phone in his hand to call, when finished, he lowers down his hand. The first and last frames of the two actions which present a person raising his hand to the head and lowering his hand down, respectively, are making the confusion, because of their reduplication in the two cases of actions within drink and call phone. Otherwise, the person performing each action is doing the same steps except that, the object handled in his hand, changes. Here, the differentiation between some actions when interacting with objects by hands becomes a key distinguishing factor for recognition. Limited by the presence of static actions like call cellphone, sit still, read book and so on in MSRDailyActivity3D, it does not allow LSTM

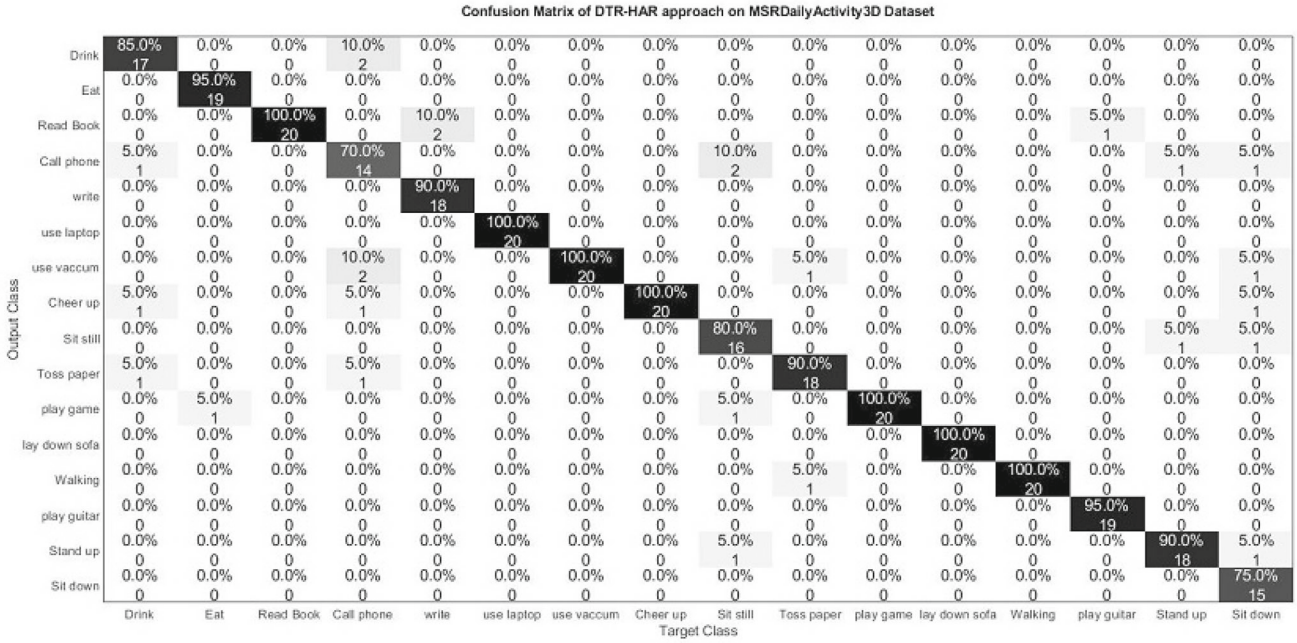


Fig. 8 Confusion matrix of DTR-HAR approach on MSRDailyActivity3D dataset

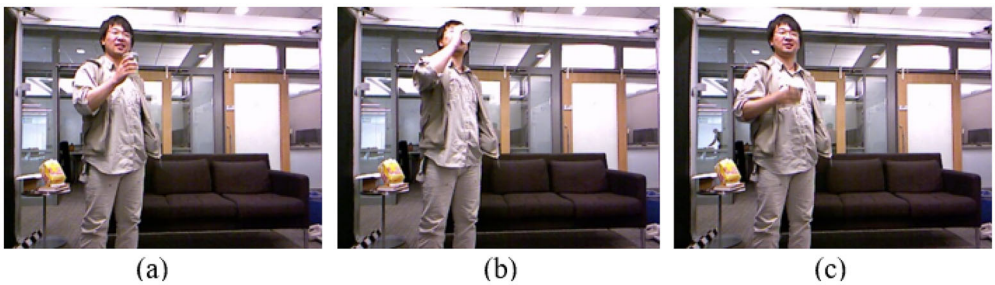


Fig. 9 Frame samples representing the action class drink. **a** A person raising his hand while holding a glass from picture set of class drink, **b** a drinking person from picture set of class drink so the glass is very close to his head, **c** a person lowering his hand after drinking from picture set of class drink

Fig. 10 Frame samples representing the action class call phone. **a** A person raising his hand while holding a phone from picture set of class call phone, **b** a call phone person from picture set of class call phone so the phone is very close to his head, **c** a person lowering his hand after calling phone from picture set of class call phone

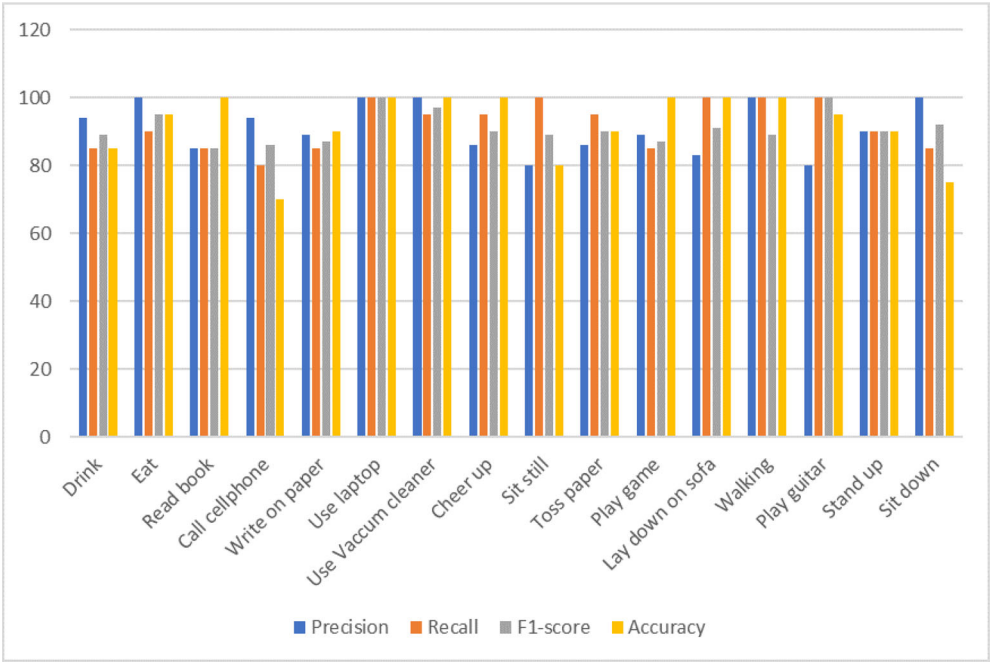


Fig. 11 Performance evaluation metrics by activity in MSRDailyActivity3D dataset

Table 5 Precision and recall of the DTR-HAR model on MSRDailyActivity3D dataset

Activity	Precision (%)	Recall (%)	F1-score (%)
Drink	94.00	85.00	89.00
Eat	100.00	90.00	95.00
Read book	85.00	85.00	85.00
Call cellphone	94.00	80.00	86.00
Write on paper	89.00	85.00	87.00
Use laptop	100.00	100.00	100.00
Use vacuum cleaner	100.00	95.00	97.00
Cheer up	86.00	95.00	90.00
Sit still	80.00	100.00	89.00
Toss paper	86.00	95.00	90.00
Play game	89.00	85.00	87.00
Lay down on sofa	83.00	100.00	91.00
Walking	100.00	100.00	89.00
Play guitar	80.00	100.00	100.00
Stand up	90.00	90.00	90.00
Sit down	100.00	85.00	92.00
Overall average	**92.00**	**92.00**	**92.00**

Bold indicates the results of our method for each evaluation metric of performance

to recognize the dynamic aspect of the activities, but we still get considerably good performance.

We repeated the evaluation of our method performance on MSRDailyActivity3D dataset. We have calculated precision, recall, F1-score and accuracy for each activity separately. Figure 11 shows the good classification of the majority of the activities. These evaluation metrics prove the efficiency of our system. The metric values shown in the Table 5 show that the DTR-HAR model achieves also good performance on MSRDailyActivity3D dataset by discriminating the activity classes of ten persons interacting in the scene. The overall achieved precision and recall are 92.00% and 92.00%, respectively. When comparing these results with those obtained on CAD-60, we noticed that the proposed system performs better with MSRDailyActivity3D data. The main cause is that the activities of CAD-60 dataset are more complex in relation with the involved poses than those contained in MSRDailyActivity3D. So, the estimation errors of

Table 6 Computational complexity comparison with state of the art

Computational complexity	AlexNet [8]	GoogLeNet [9]	ResNet50 [7]	Our Model
Spatial (millions)	60M	4M	0.85M	**1.7M**
Temporal (flops)	1.5×10^9	1.5×10^9	3.8×10^9	$\mathbf{7.6 \times 10^9}$

Bold indicates the results of our method for each evaluation metric of performance

a single pose accumulate, causing less reliable recognition process.

Computational complexity Introducing additional layers and nodes to a neural network, makes it deeper. This is a critical approach to improve its performance and in return increase its computational complexity. Consequently, it is crucial and interesting to resolve the problem of high computation cost to realize real time and reliable human activity recognition by deep learning models. Table 6 shows that our network model converges well. It has fewer parameters than other deep networks.

6 Conclusion

In this work, we proposed a deep temporal residual neural network architecture which is used to model spatiotemporal sequence video information in order to enhance the performance facing the human activity recognition challenge. A deep convolutional neural network-based residual network model is used to extract discriminative visual spatial features and a LSTM neural network to deal with the long-term temporal dependencies whose order has a great influence on the context of the performed action. In fact, we have exploited residual spatial features to extract temporal features by applying activation to the last fully connected layer of the trained convolutional residual network. Our approach was validated on two benchmark datasets and have obtained competitive results to the state-of-the-art performances. In the future, first, we intend to incorporate different modalities such as optical flow and depth maps to seek for the best architecture to fuse them. Second, we want to combine with hand gesture to better distinguish the object handled during the implementation of the action.

Compliance with ethical standards

Conflict of interest Hend Basly, Wael Ouarda, Fatma Ezahra Sayadi, Bouraoui Ouni and Adel M. Alimi declare that they have no conflict of interest.

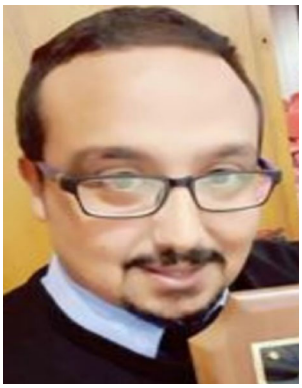
References

1. Zhou, T., Wang, W., Qi, S., Ling, H., Shen, J.: Cascaded human–object interaction recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4263–4272. IEEE (2020)
2. Wang, W., Zhu, H., Dai, J., Pang, Y., Shen, J., Shao, L.: Hierarchical human parsing with typed part-relation reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8929–8939. IEEE (2020)
3. Li, T., Liang, Z., Zhao, S., Gong, J., Shen, J.: Self-learning with rectification strategy for human parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9263–9272. IEEE (2020)
4. Qi, S., Wang, W., Jia, B., Shen, J.,Zhu, S.C.: Learning human–object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 401–417. Springer (2018)
5. Yilmaz, A., Shah, M.: Actions sketch: a novel action representation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 984–989. IEEE (2005)
6. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision (ICCV'05), pp. 1395–1402. IEEE (2005)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Las Vegas, NV (2016)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . , Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
11. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732. IEEE, Columbus, OH (2014)
12. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: 15th ACM International Conference on Multimedia, pp. 357–360. ACM, Augsburg, Germany (2007)
13. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: 19th British Machine Vision Conference (BMVC), pp. 1–10. Leeds, United Kingdom (2008)
14. Oreifej, O., Liu, Z.: 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. IEEE, Portland, OR (2013)
15. Asadi-Aghbolaghi, M., Kasaei, S.: Supervised spatio-temporal kernel descriptor for human action recognition from RGB-depth videos. Multimed. Tools Appl. **77**(11), 14115–14135 (2018). https://doi.org/10.1007/s11042-017-5017-y
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94

17. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008). https://doi.org/10.1016/j.cviu.2007.09.014
18. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: European Conference on Computer Vision, pp. 650–663. Springer, Berlin, Heidelberg (2008)
19. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: 12th International Conference on Computer Vision, pp. 128–135. IEEE, Kyoto (2009)
20. Zhang, M., Sawchuk, A.A.: Human daily activity recognition with sparse representation using wearable sensors. IEEE J. Biomed. Health. Inf. **17**(3), 553–560 (2013). https://doi.org/10.1109/JBHI.2013.2253613
21. Liu, C., Ying, J., Yang, H., Hu, X., Liu, J.: Improved human action recognition approach based on two-stream convolutional neural network model. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01868-8
22. Li, X., Shen, H., Zhang, L., Zhang, H., Yuan, Q., Yang, G.: Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. IEEE Trans. Geosci. Remote Sens. **52**(11), 7086–7098 (2014). https://doi.org/10.1109/TGRS.2014.2307354
23. Dong, X., Shen, J., Wu, D., Guo, K., Jin, X., Porikli, F.: Quadruplet network with one-shot learning for fast visual object tracking. IEEE Trans. Image Process. **28**(7), 3516–3527 (2019). https://doi.org/10.1109/TIP.2019.2898567
24. Liang, Z., Shen, J.: Local semantic Siamese networks for fast tracking. IEEE Trans. Image Process. **29**, 3351–3364 (2019). https://doi.org/10.1109/TIP.2019.2959256
25. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Trans. Image Process. **27**(1), 38–49 (2017). https://doi.org/10.1109/TIP.2017.2754941
26. Lai, Q., Wang, W., Sun, H., Shen, J.: Video saliency prediction using spatiotemporal residual attentive networks. IEEE Trans. Image Process. **29**, 1113–1126 (2019). https://doi.org/10.1109/TIP.2019.2936112
27. Wang, W., Shen, J., Ling, H.: A deep network solution for attention and aesthetics aware photo cropping. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1531–1544 (2018). https://doi.org/10.1109/TPAMI.2018.2840724
28. Kuanar, S., Athitsos, V., Pradhan, N., Mishra, A., Rao, K.R.: Cognitive analysis of working memory load from EEG, by a deep recurrent neural network. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2576–2580. IEEE SigPort (2018)
29. Kuanar, S., Athitsos, V., Mahapatra, D., Rao, K.R., Akhtar, Z., Dasgupta, D.: Low dose abdominal CT image reconstruction: an unsupervised learning based approach. In: IEEE International Conference on Image Processing (ICIP), pp. 1351–1355. IEEE (2019)
30. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
31. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576. MIT Press, Cambridge, MA (2014)
32. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497. IEEE Computer Society, USA (2015)
33. Bhattacharya, S., Nurmi, P., Hammerla, N., Plötz, T.: Using unlabeled data in a sparse-coding framework for human activity recognition. Pervasive Mob. Comput. **15**, 242–262 (2014). https://doi.org/10.1016/j.pmcj.2014.05.006
34. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization (2014). arXiv preprint arXiv:1409.2329
35. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
36. . Veeriah, V., Zhuang, N., Qi, G.J.: Differential recurrent neural networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4041–4049. IEEE Computer Society, USA (2015)
37. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694–4702. Boston, MA (2015)
38. Wu, Z., Wang, X., Jiang, Y.G., Ye, H., Xue, X.: Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 461–470. ACM, New York, NY (2015)
39. Kim, J.H., Hong, G.S., Kim, B.G., Dogra, D.P.: deepGesture: deep learning-based gesture recognition scheme using motion sensors. Displays **55**, 38–45 (2018). https://doi.org/10.1016/j.displa.2018.08.001
40. Madhuranga, D., Madushan, R., Siriwardane, C., Gunasekera, K.: Real-time multimodal ADL recognition using convolution neural networks. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01864-y
41. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using LSTMs. In: International Conference on Machine Learning, pp. 843–852. JMLR.org (2015)
42. Ercolano, G., Riccio, D., Rossi, S.: Two deep approaches for ADL recognition: a multi-scale LSTM and a CNN-LSTM with a 3D matrix skeleton representation. In: 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 877-882. IEEE, Lisbon (2017)
43. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W.: Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access **6**, 1155–1166 (2017). https://doi.org/10.1109/ACCESS.2017.2778011
44. Ma, C.Y., Chen, M.H., Kira, Z., AlRegib, G.: TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition. Signal Process. Image Commun. **71**, 76–87 (2019). https://doi.org/10.1016/j.image.2018.09.003
45. Zhao, R., Ali, H., Van der Smagt, P.: Two-stream RNN/CNN for action recognition in 3D videos. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4260–4267. IEEE, Vancouver, BC (2017)
46. Ding, L., Fang, W., Luo, H., Love, P.E., Zhong, B., Ouyang, X.: A deep hybrid learning model to detect unsafe behavior: integrating convolution neural networks and long short-term memory. Autom. Constr. **86**, 118–124 (2018). https://doi.org/10.1016/j.autcon.2017.11.002
47. Baradel, F., Wolf, C., Mille, J.: Human activity recognition with pose-driven attention to RGB. In: 29th British Machine Vision Conference, pp. 1–14. Newcastle, United Kingdom (2018)
48. Arif, S., Wang, J., Ul Hassan, T., Fei, Z.: 3D-CNN-based fused feature maps with LSTM applied to action recognition. Future Internet **11**(2), 42 (2019). https://doi.org/10.3390/fi11020042
49. Das, S., Koperski, M., Bremond, F., Francesca, G.: Deep-temporal LSTM for daily living action recognition. In: 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE, Auckland, New Zealand (2018)
50. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Berg, A.C.: Imagenet large scale visual recognition challenge. Int.

J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
51. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: International Conference on Machine Learning, pp. 1310–1318. JMLR.org, Atlanta, GA (2013)
52. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from RGBD images. In: IEEE International Conference on Robotics and Automation, pp. 842–849. IEEE, Saint Paul, MN (2012)
53. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1290–1297. IEEE, USA (2012)
54. Khaire, P., Kumar, P., Imran, J.: Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognit. Lett. **115**, 107–116 (2018). https://doi.org/10.1016/j.patrec.2018.04.035
55. Zhu, Y., Chen, W., Guo, G.: Evaluating spatiotemporal interest point features for depth-based action recognition. Image Vis. Comput. **32**(8), 453–464 (2014). https://doi.org/10.1016/j.imavis.2014.04.005
56. Ni, B., Pei, Y., Moulin, P., Yan, S.: Multilevel depth and image fusion for human activity detection. IEEE Trans. Cybern. **43**(5), 1383–1394 (2013). https://doi.org/10.1109/TCYB.2013.2276433
57. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from RGB-D videos. Int. J. Robot. Res. **32**(8), 951–970 (2013). https://doi.org/10.1177/0278364913478446
58. Koperski, M., Bremond, F.: Modeling spatial layout of features for real world scenario RGB-D action recognition. In: 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 44–50. IEEE (2016)
59. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for RGB-D activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5344–5352. IEEE (2015)
60. Srihari, D., Kishore, P.V.V., Kumar, E.K., Kumar, D.A., Kumar, M.T.K., Prasad, M.V.D., Prasad, C.R.: A four-stream ConvNet based on spatial and depth flow for human action classification using RGB-D data. Multimed. Tools Appl. (2020). https://doi.org/10.1007/s11042-019-08588-9
61. Nunez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S., Velez, J.F.: Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit. **76**, 80–94 (2018). https://doi.org/10.1016/J.PATCOG.2017.10.033
62. Luo, J., Wang, W., Qi, H.: Spatio-temporal feature extraction and representation for RGB-D human action recognition. Pattern Recognit. Lett. **50**, 139–148 (2014). https://doi.org/10.1016/j.patrec.2014.03.024
63. Baradel, F., Wolf, C., Mille, J.: Human activity recognition with pose-driven attention to RGB. In: 29th British Machine Vision Conference, pp. 1–14. Newcastle, United Kingdom (2018)
64. Zhou, Y., Ni, B., Hong, R., Wang, M., Tian, Q.: Interaction part mining: a mid-level approach for fine-grained action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3323–3331. IEEE (2015)
65. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision, pp. 20–36. Springer, Cham (2016)
66. Das, S., Koperski, M., Bremond, F., Francesca, G.: Action recognition based on a mixture of RGB and depth based skeleton. In: 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE, Lecce (2017)
67. Liu, L., Shao, L.: Learning discriminative representations from RGB-D video data. In: International Joint Conference on Artificial Intelligence (2013)
68. Kong, Y., Fu, Y.: Bilinear heterogeneous information machine for RGB-D action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1054–1062 (2015)

Hend Basly received her Master degree in Computer Science from Higher Institute of Applied Sciences and Technology (ISSAT), University of Sousse, in 2014. She is currently a PhD student at Networked Objects, Control, and Communication Systems (NOCCS), Laboratory at the National School of Engineers of Sousse (ENISO). Her research interests are focused on deep learning neural networks, video surveillance, and human activity recognition.



Wael Ouarda obtained his PhD in Computer Science from the Research Groups in Intelligent Machines (ReGIM-Lab) at the National School of Engineers of Sfax (ENIS) in 2017. He is currently a Professor-Researcher in Computer Science at the University of Sfax and quality manager at the ReGIM-Lab since 2016. Dr. Wael OUARDA has more than 30 publications in prestigious conferences and journals in the field of Artificial Intelligence. His research focuses on image analysis (biometric, medical, and social) with deep learning techniques of neural networks for the representation of features and for classification. Dr. Wael OUARDA has been an active IEEE member since 2012.

Fatma Ezahra Sayadi is currently an Associate Professor at the National Engineering School of Sousse. She received her PhD Degree in Micro-electronics from Faculty of Science of Monastir, Tunisia, in collaboration with the LESTER Laboratory, University of South Brittany Lorient FRANCE, in 2006. She is currently a member of the Laboratory of Electronics & Microelectronics. Her research includes image and video processing in graphics processor, motion tracking and pattern recognition, circuit and system design.

Bouraoui Ouni is currently a Professor at the National Engineering School of Sousse. He has obtained his PhD entitled âAŸSynthesis and temporal partitioning for reconfigurable systemsâAZ in 2008 from the Faculty of Sciences at Monastir. He obtained his university habilitation entitled "Optimisation algorithm for reconfigurable architectures" in 2012. Hence, his research interests cover: models, methods, tools, and architectures for reconfigurable computing; simulation, debugging, synthesis, verification, and test of reconfigurable systems; field-programmable gate arrays and other reconfigurable technologies; algorithms implemented on reconfigurable hardware; hardware/software codesign and cosimulation with reconfigurable hardware; and high-performance reconfigurable computing.

Adel M. Alimi graduated in Electrical Engineering 1990 and obtained a PhD and then an HDR both in Electrical and Computer Engineering in 1995 and 2000, respectively. He is now professor in Electrical and Computer Engineering at the University of Sfax. His research interest includes applications of intelligent methods (neural networks, fuzzy logic, evolutionary algorithms) to pattern recognition, robotic systems, vision systems, and industrial processes. He focuses his research on intelligent pattern recognition, learning, analysis, and intelligent control of large-scale complex systems. He is associate editor and member of the editorial board of many international scientific journals (e.g., "IEEE Trans. Fuzzy Systems," "NeuroComputing," "Neural Processing Letters," "International Journal of Image and Graphics," "Neural Computing and Applications," "International Journal of Robotics and Automation," "International Journal of Systems Science," etc.). He was guest editor of several special issues of international journals (e.g., Fuzzy Sets & Systems, Soft Computing, Journal of Decision Systems, Integrated Computer Aided Engineering, Systems Analysis Modelling and Simulations). He is the Founder and Chair of many IEEE Chapter in Tunisia section, he is IEEE Sfax Subsection Chair (2011), IEEE ENIS Student Branch Counselor (2011), IEEE Systems, Man, and Cybernetics Society Tunisia Chapter Chair (2011), IEEE Computer Society Tunisia Chapter Chair (2011), and he is also Expert evaluator for the European Agency for Research. He was the general chairman of the International Conference on Machine Intelligence ACIDCA-ICMI'2005 & 2000. He is an IEEE senior member.