An Efficient Human Activity Recognition Technique Based on Deep Learning

A. Khelalef[a,*], F. Ababsa[b,**], and N. Benoudjit [a,***]

[a] *Laboratoire d'Automatique Avancée et d'Analyse des Systèmes (LAAAS), University Batna-2, Batna, Algeria*
[b] *Ecole Nationale d'Arts et Métiers, Institut image-Le2i, Paris, France*
** e-mail: Khelalef_aziz@yahoo.fr*
*** e-mail: fakhreddine.ababsa@ensam.eu*
**** e-mail: n.benoudjit@univ-batna2.dz*

Abstract—In this paper, we present a new deep learning-based human activity recognition technique. First, we track and extract human body from each frame of the video stream. Next, we abstract human silhouettes and use them to create binary space-time maps (BSTMs) which summarize human activity within a defined time interval. Finally, we use convolutional neural network (CNN) to extract features from BSTMs and classify the activities. To evaluate our approach, we carried out several tests using three public datasets: Weizmann, Keck Gesture and KTH Database. Experimental results show that our technique outperforms conventional state-of-the-art methods in term of recognition accuracy and provides comparable performance against recent deep learning techniques. It's simple to implement, requires less computing power, and can be used for multi-subject activity recognition.

Keywords: human activity recognition, deep learning, convolutional neural network (CNN), features extraction, classification

DOI: 10.1134/S1054661819040084

1. INTRODUCTION

Nowadays, Human activity recognition is one of the most important fields in computer vision research; it has large applications in industrial and common life routines; it is used in video surveillance, human-machine interaction, monitoring systems, virtual reality and many other applications.

The challenge in human activity recognition is to efficiently recognize various actions in complex situations, to provide a high accuracy recognition rate, and to simplify implementation in real time application while using less computing power.

View-based human activity recognition techniques use space-time information in the video stream to recognize human actions by extracting specific features. Generally, it consists of two steps: (1) pre-processing and features extraction during which the aim is to prepare the data for the second step by applying different operations like resizing, background subtraction, extracting silhouettes or skeletons, applying transforms such as DCT (Discrete Cosine Transform) or FT (Fourier Transform), (2) features extraction step consisting of features calculation from the pre-processed data. Features extraction techniques can be classified in three categories: Methods using global features, Local features and Body modeling techniques.

Many View-based human activity recognition methods were proposed in the literature. Earlier works developed several methods using global features [1]. In [2] Blank et al. used silhouettes to create a space-time volume from which space-time saliency, shape structure and orientation are extracted. In [3] Dollar et al. proposed to extract the local region of interest from space-temporal volume to create distinguishable features used for recognition. In [4] Kumari and Mitra proposed a transform-based technique by using discrete Fourier transforms (DFTs) of small image blocks as features. Furthermore, in [5] Tasweer et al. used motion history image (HMI) to extract features by using a blocked discrete cosine transform (DCT). In [6] Hafiz et al. used Fourier transform domain of frames to extract spectral features and principal component analysis (PCA) to reduce the features dimension.

Local features are also widely used in human activity recognition, in [7] Lowe introduced the SIFT descriptor (scale invariant feature transform) which enable the extraction of a robust local features invariant to image scaling, translation and rotation. In [8] Dalal and al. proposed the oriented gradient descriptors (HOG) for human activity recognition, by calculating the gradient orientation in portions of the image as features for recognition. In [9] Lu and Little pro-

Received March 8, 2019; revised July 11, 2019;
accepted July 15, 2019

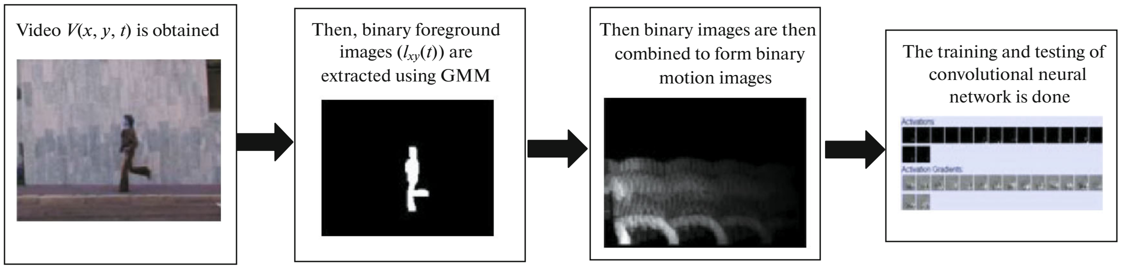

Video $V(x, y, t)$ is obtained
Then, binary foreground images ($I_{xy}(t)$) are extracted using GMM
Then binary images are then combined to form binary motion images
The training and testing of convolutional neural network is done
Activations
Activation Gradients

Fig. 1. (Color online) Overview of the proposed technique in [22].

posed the PCA-HOG which is an improvement of the HOG descriptor by using (principal component analysis) PCA in order to create local descriptors invariant to illumination, pose and viewpoint. In [10, 11]. Matti Pietikäinen et al. introduced Local Binary Patterns (LBP) for texture classification, it consists of extracting histograms of quantized local binary patterns in a local region of the image. In [12] Lin et al. proposed a nonparametric weighted feature extraction (NWFE) approach by using PCA (principal component analysis) and K-means clustering to build histogram vectors from pose contour.

Body modeling human activity recognition techniques are also widely used; here the human body is modeled to be tracked and recognized. In [13] Nakazawa et al. represent and track the human body by using an ellipse. In [14] Iwasawa et al. proposed to create human skeleton models using sticks. In [15] Huo et al. proposed to model the human head, shoulder and upper-body for recognition, in [16] Sedai et al. used a 3D human body modeling of 10 body parts (torso, head, arms, and legs...).

Recently deep learning can be considered a revolutionary tool in computer vision research. The capability of convolutional neural network to create distinguishable features directly from the input images using multiple hidden layers makes the introduction of this tool quite interesting in the domain of human activity recognition. More recently, most of the applications of deep learning in human activity recognition has been relying on the use of wearable sensors [17–21]. However, research using view-based approaches remains scarce.

In this paper, we present a new deep learning–based human activity recognition technique. The objective is to recognize human activities in a video stream using extracted binary space-time maps (BSTMs) as the input of the Convolutional neural network (CNN). The main contributions of our paper are summarized as follows:

- We propose a simple deep learning-based method consisting of two steps: (1) binary space-time maps (BSTMs) extraction from silhouettes of segmented and centred human body, (2) features extraction and action classification using convolutional neural network (CNN).
- The proposed technique offers the capability to recognize multiple actions in the same video frame because the BSTMs are extracted only from the silhouettes of segmented human body.
- Experimental investigations using multiple benchmark databases (Weizmann, Keck Gesture and KTH databases) showing that our technique is efficient and outperforms conventional human activity recognition methods and gives comparable performance against recent deep learning-based techniques.

This paper is organized as follow, in section two, we present state-of-the-art of deep learning-based techniques. Next, we give a brief introduction to CNN. We present the proposed method in section four. Experimental results will be given in the section that follows. Finally, section six contains concluding remarks and future perspectives.

2. DEEP LEARNING-BASED TECHNIQUES – RELATED WORK

Deep learning capability to self-extract distinguishable features yields to open a new era in human activity recognition field; in this section we review recent approaches.

In [22] Tushar D. et al. proposed a deep learning-based technique using binary motion image (BMI), the authors used Gaussian Mixture Models (GMM) to subtract binary backgrounds used to create BMIs (Fig. 1), and three (3) CNN layers to extract features and classify activities. BMIs are extracted from the frames, which make the use of this approach impossible for multi-human recognition.

Moez B. et al. proposed in [23], a two-steps neural recognition method (Fig. 2) using an extension of convolutional neural network to 3D to learn spatial-temporal features. The authors proposed to extract the features using 10 layers of CNN (input layer, two combinations of convolution/rectification/sub-sampling layers, a third convolution layer and two neuron layers). For the recognition step, they proposed to use a recurrent neural network classifier (Long Short-Term

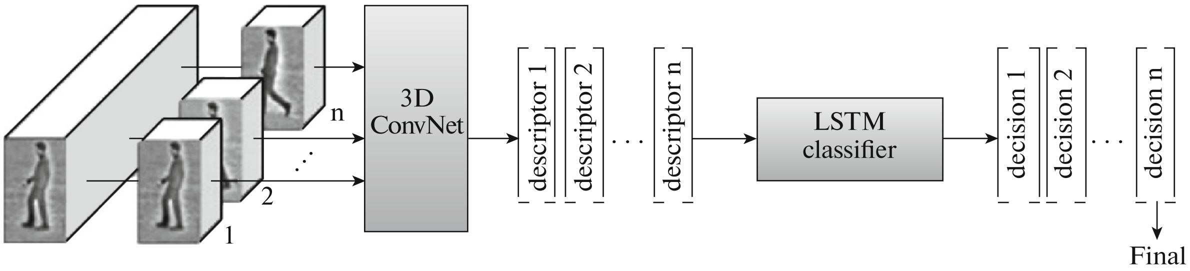


Fig. 2. A two steps neural recognition scheme from [23].

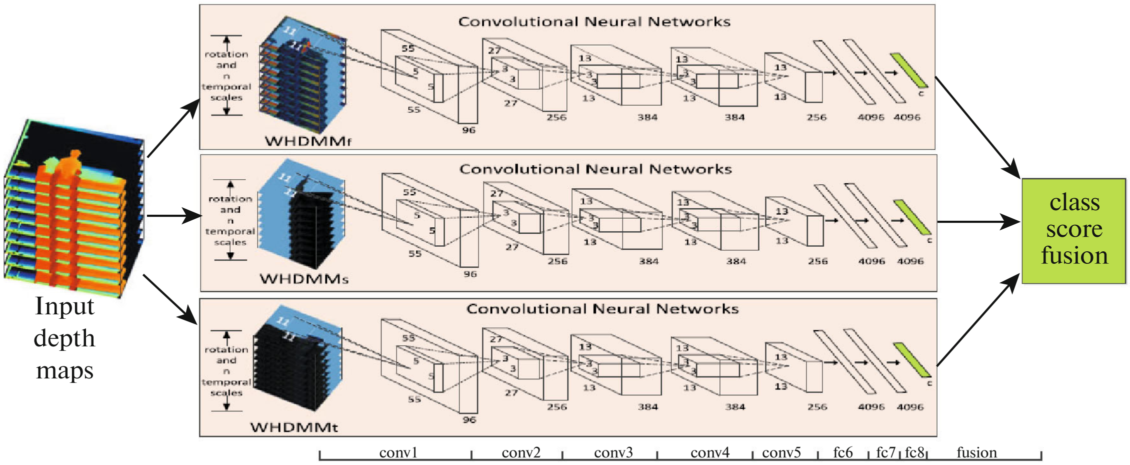


Fig. 3. (Color online) Proposed method in [25].

Memory (LSTM) classifier) by taking advantage of the temporal evolution of the features.

In [24] Pichao Wang et al. used a weighted hierarchical depth motion map (WHDMM) and three channel deep CNN for human activity recognition. Here, the authors proposed to feed three separate ConvNets using WHDMMs constructed by the projection the 3D points of depth images to three orthogonal planes, the final classification decision is obtained by the fusion of the three ConvNets.

In [25] Zuxuan W. et al. constructed a hybrid method for video classification by extracting two types of features from spatial frames (raw frames) and short-term stacked motion optical flows using convolutional neural network (Fig. 3). These features are used to feed two separate LSTM networks for fusion and classification.

The authors in [26] proposed a human activity recognition approach using depth images, from which they proposed to extract three derived images (Fig. 4): Average depth image (ADI), Motion history image (MHI) and depth difference image and used a deep belief network (DBN) using a Restricted Boltzmann Machine (RBM).

Andrej K. et al. in [27], proposed a multi-resolution convolutional neural network approach (Fig. 5), here the authors used two ConvNet channels, the first channel is fed by a context stream representing a low-resolution image; the second one is fed by a fovea stream representing a high resolution centred image. The two channels converge towards two fully connected layers.

In [28] Simonyan et al. presented a two-stream architecture for video classification (Fig. 6), the authors proposed to use a spatial stream ConvNet using raw video frames to carry information about the objects and the general spatial information in the scenes, and a Temporal stream ConvNet using optical flow from multiple input video frames. Classification is done using two fusion methods: the average of the two stream scores or by using a multiclass SVM on Softmax scores.

In this paper, we present a simple and efficient human activity recognition technique using deep-learning. Our work is inspired by papers [22, 29] where authors proposed to use Motion Energy Images (MEI) and Motion History images (MHI) for human activity recognition, however, unlike the aforemen-

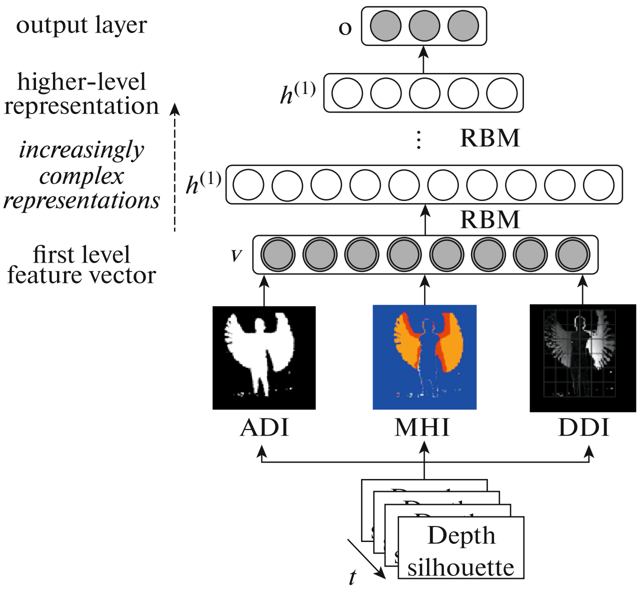

output layer
o
higher-level representation
$h^{(1)}$
RBM
increasingly complex representations
$h^{(1)}$
RBM
first level feature vector
v
ADI
MHI
DDI
Depth silhouette
t

Fig. 4. (Color online) An overview of the proposed approach from [26].

tioned works, our technique operates only on the segmented human body which make it suitable for multi-human activity recognition. In the next section, we will give a brief introduction to CNN.

3. CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN were proposed for the first time by Lecun Y. and Bengio Y. in [30]. It has the capability to extract feature maps directly from the input images and classify theme into many categories by using successive combinations of convolution/sub-sampling hidden layers and being invariant for shift and distortions.

For a better understanding of the convolutional neural network architecture, we take the example of convolutional neural network presented in [30] for handwriting recognition (Fig. 7).

The convolutional neural network is composed of multiple hidden layers, those layers are a successive combination of convolution and sub-sampling operations. Each convolution layer is composed of multiple feature maps, and each convolution/sub-sampling combination has the same number of feature maps [26].

The first layer of the CNN is the input layer; it has the same dimension as the input images, the first hidden layer is obtained by the convolution of the input layer by the kernel [22], the second hidden layer is obtained by performing a 2 by 2 averaging and subsampling. The next hidden layers are found in the same way by using a successive alternation of convolutions and sub-sampling, each unit of a layer is fully connected to the units of the previous layer. The number of feature maps increases and the resolution decreases at each convolution/sub-sampling combination [30]. The last layer is the classification layer. It contains the last feature map and its dimension is the number of classes to recognize.

4. PROPOSED TECHNIQUE: BSTM DEEP LEARNING RECOGNITION

The proposed technique consists of two steps divided into five processes: Human detection and tracking, human pose extraction, silhouettes extraction, binary space-time map (BSTM) calculation and deep-learning recognition (Fig. 8).

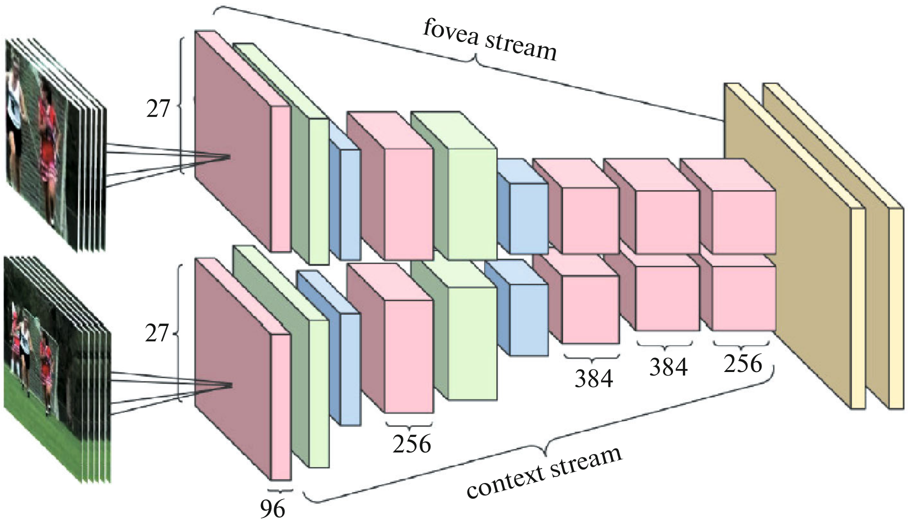


Fig. 5. (Color online) Multiresolution CNN in [27].

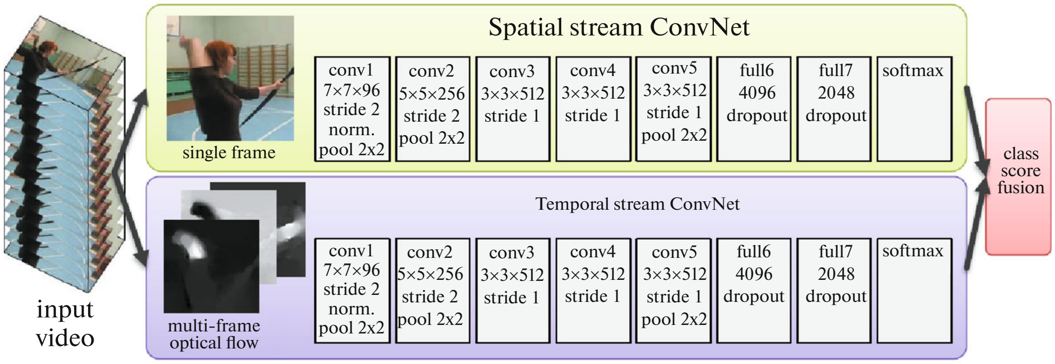


Fig. 6. (Color online) Two-stream architecture used in [28].

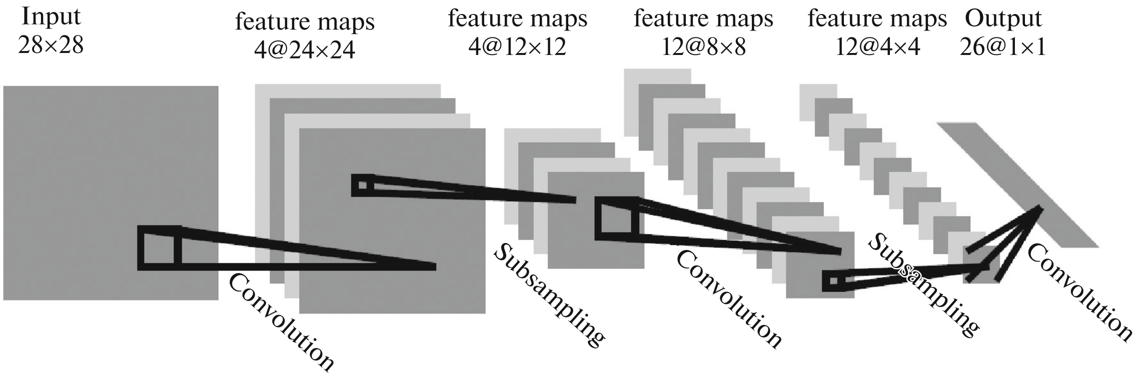


Fig. 7. Example of the convolutional neural network for Handwriting recognition proposed by Lecun Y. et al. in [30].

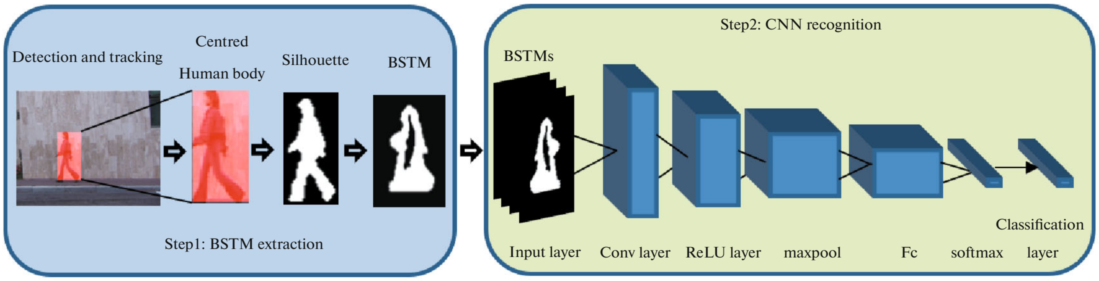


Fig. 8. (Color online) Overview of the proposed deep learning-based technique.

The first step in the proposed scheme is human detection and tracking, for that we implemented a simple background extraction algorithm to detect and track the human body using the variations of intensity of the foreground images. Then for each frame, human body is segmented, and silhouettes extracted using Otsu's image segmentation algorithm [31] by thresholding the images using an optimum threshold (thr) that minimize the weighted within-class variance [31].

The results of Otsu's segmentation algorithm are binary images:

If we denote $g(x, y)$ is the threshold version of the original grey scale image $f(x, y)$ using the threshold Thr then [31]:

$$g(x, y) = \begin{cases} 1 & \text{if} \quad f(x, y) \geq Thr \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

Binary space-time maps (BSTM) are like Motion Energy Images (MEI), it is a binary template created

Table 1. BSTM extraction algorithm

```
Input: Subject silhouettes $g(x, y)$
Output: Binary space-time map $h(x, y)$

PROCESS:
  For i = 1:T
    $h(x, y) = \sum abs(g_{ti+1}(x, y) - g_{t1}(x, y))$
  End
  If $h(x, y) \neq 0$
    $h(x, y) = 1$
  Else
    $h(x, y) = 0$
  End
End
```

from centred and segmented human silhouettes of each frame by using our proposed algorithm shown in table 1 below:

Let denote:

$g(x, y)$: the extracted silhouettes,

$h(x, y)$: Binary space-time map,

T: The number of **frames.**

The extracted BSTMs contain the space-time information of the human action in a lap of time; the quality of the BSTMs depends on the quality of extracted silhouettes and on the number of frames used. Our investigations show that generally if the quality of the used silhouettes is acceptable, sixteen Frames are sufficient to create distinguishable BSTMs. An example of extracted BSTMs using Keck Gesture Dataset is shown below (Fig. 9).

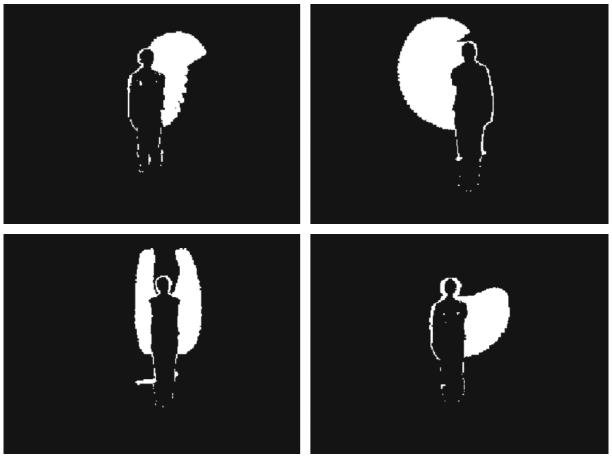


Fig. 9. Samples of binary space-time maps (BSTM) using Keck Gesture dataset.

Unlike the techniques proposed in [22, 23], our proposed Method offers the ability to track and recognize multiple subjects in the same frame because we calculate the BSTMs only from the extracted human body not from the entire frames.

And because the constructed BSTMs are simple images that contain the binary space-time information of the human body in a lap of the time, our technique is simpler, faster and does not require many computing powers.

The next stage in our proposed deep learning-based method is features extraction and classification. Here we used the capability of deep learning (CNN in our case) to automatically extract and classify the input data.

We trained the convolutional neural network using our proposed binary space-time maps (BSTMs) extracted from the video frames.

The architecture of the proposed seven layers CNN is as follow:

- The input layer: has the same dimension as the input BSTMs, an example of extracted BSTM is shown in (Fig. 9).

- Convolutional layer: the objective of the convolutional layer is feature extraction in sub regions that depend on the filter size. Activations of the convolutional layer shows the area in the convolutional layer that activate on the input BSTM image (Fig. 10).

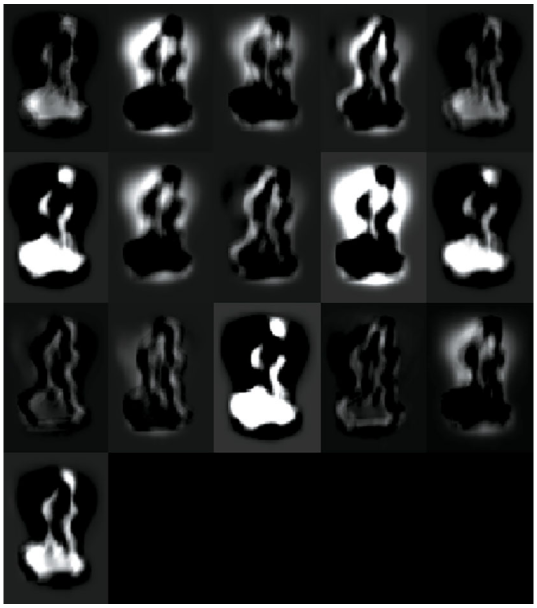


Fig. 10. Example of activation of convolutional neural network to the running activity.

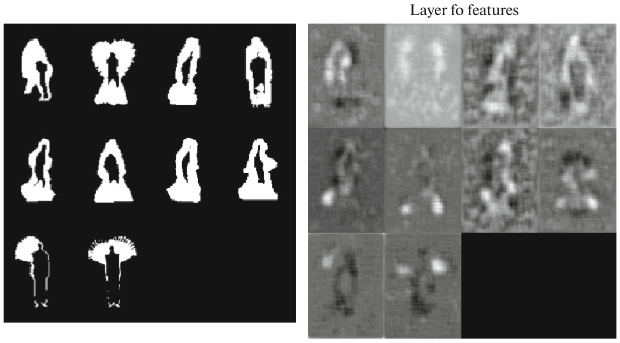


Fig. 11. Samples of extracted (fully connected layer) features using CNN (Weizmann database): (Left) extracted BSTMS, (right) features from CNN (last fully connected layer).

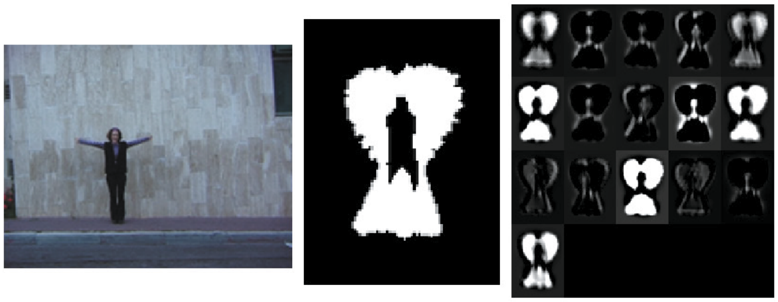


Fig. 12. (Color online) Sample from Weizmann dataset: (left) Wave2, (Middle) extracted BSTM, (right) activations of the first convolutional layer.

In this paper, we propose to use one convolutional layer for all datasets, table 2 shows the size of the filters used for each d**atabase.**

Rectified linear unit layer (ReLU): this layer performs a threshold operation and set to zero any input less than zeros.

- Maxpool layer: it is used to reduce the number of parameters fed to feed to the next layer by performing a sub-sampling operation.
- Fully connected layer: it contains the extracted features learned by the earlier layers, Fig. 11 shows the BSTMs and the extracted features in the fully connected layer from Weizmann database.
- Softmax layer: this is a vector calculated using the Softmax activation function, which is the generalization of the sigmoid function.
- Classification layer: this is the classification layer, it returns the final class of the constructed BSTM using the results of the Softmax layer.

5. EXPERIMENTAL RESULTS

To evaluate and validate the performance and efficiency of the proposed method, we carried out several tests using three benchmark datasets: Weizmann database [32], Keck Gesture Dataset [33] and KTH dataset [34]. We carried out several tests using different sets of learning and testing.

To enable comparison against state-of-the-art methods we used two standard evaluation criteria: recognition rate and one-versus-rest ROC curve.

The recognition rate used is defined by:

$$\begin{aligned}&\mathit{Recognition\ rate}\\&=\frac{\mathit{Number\ of\ good\ classified\ actions}}{\mathit{Total\ Number\ of\ actions}}.\end{aligned} \tag{2}$$

Table 2. Convolutional layer filter's size used for each database

Database	Convolutional layer filter's size	Number of filters	Stride
Weizmann	[10 × 10]	16	1
Keck Gesture Dataset	[5 × 5]	10	1
KTH	[3 × 3]	10	1

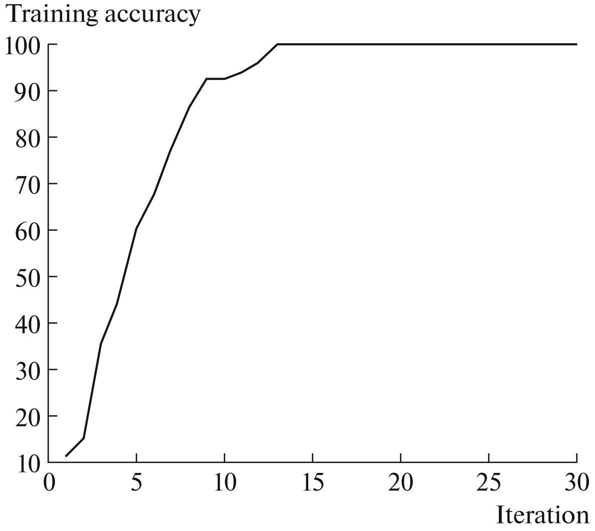


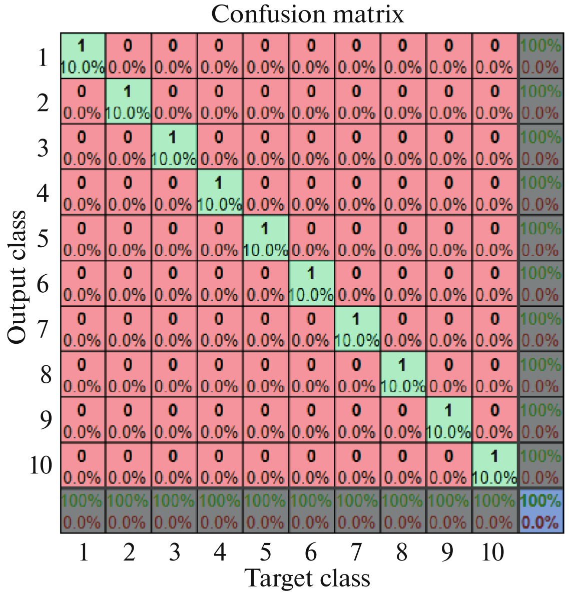
Fig. 13. Training accuracy rate against the current iteration using Weizmann database.

5.1. Weizmann Database

Weizmann database [32] consists of 10 activities (Bend, Jack, Jump, Pjump, Run, Side, Skip, Walk, Wave1 and Wave2) performed by nine persons i.e. 90 actions, using a simple background and a fixed camera. Figure 12 shows an example of an action (wave2) from Weizmann database, extracted BSTMs and the activations of convolutional layer of the CNN. To allow a direct comparison to recent works, we trained our ConvNet using videos from eight actors and tested on the remaining one.

Table 3. Human activity recognition rates obtained in literature and in our approach using Weizmann database

Method	Accuracy
Boiman and Irani 2006 [35]	97.5% (9 actions)
Scovanner et al. 2007 [36]	82.6% (10 actions)
Wang and Suter 2007 [37]	97.8% (10 actions)
Kellokumpu et al 2008 [38]	97.8% (10 actions)
Kellokumpu et al. 2009 [39]	98.7% (9 actions)
Hafiz Imtiaz et al. 2015 [6]	100% (10 actions)
Tasweer et al. 2015 [5]	92.25% (10 actions)
Aziz et al. 2016 [40]	99.03% (10 actions)
Tushar et al. 2015 [22]	100% (5 actions)
Our approach	**100% (10 actions)**
Our approach	**100% (9 actions)**



	1	2	3	4	5	6	7	8	9	10	
1	1 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	1 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	1 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	1 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 10.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 10.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 10.0%	0 0.0%	0 0.0%	100% 0.0%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 10.0%	0 0.0%	100% 0.0%
10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 10.0%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Fig. 14. (Color online) Confusion matrix using Weizmann database.

Comparative study using Weizmann dataset (Table 3) shows that our proposed method outperforms conventional state-of-the-art methods and deep-learning based methods. (in [22] the authors used only 5 activities) and achieved rapidly the highest training accuracy rate (Fig. 13) even when using 10 activities (Fig. 14).

ROC curve in (Fig. 15), shows the efficiency of our classifier and the effectiveness of the proposed BSTMs.

5.2. Keck Gesture Dataset

We carried out several tests using Keck Gesture Dataset [33]. The test consists of 14 activities (Turn left, turn right, Att-left, Att-right, Att-both, Stop left, Stop right, Stop both, Flap, Start, come near, Close Dis, speed up, go back) performed by three individuals, i.e. 42 video sequences. Figure 16 shows samples of frames from Keck Gesture Dataset, their extracted BSTMs and the activations of convolutional layer of the ConvNet. In the experimental setup, we used the actions performed by two individuals for training, and we tested using the third one.

The experimental results of the proposed method (Table 4) against state-of-the-art in [33] shows that the outcome of our proposed method provides a perfect recognition rate when using Keck Gesture database, and that it outperforms the original approach.

Figure 17 illustrates the training accuracy when using Keck Gesture database. Thus, our proposed CNN architecture achieves 100% training accuracy significantly fast. The efficiency of our classifier is

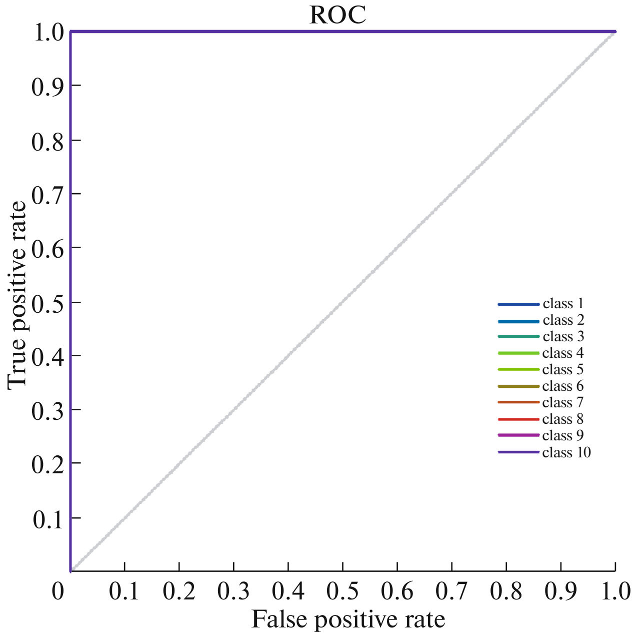


Fig. 15. (Color online) ROC curve using Weizmann database.

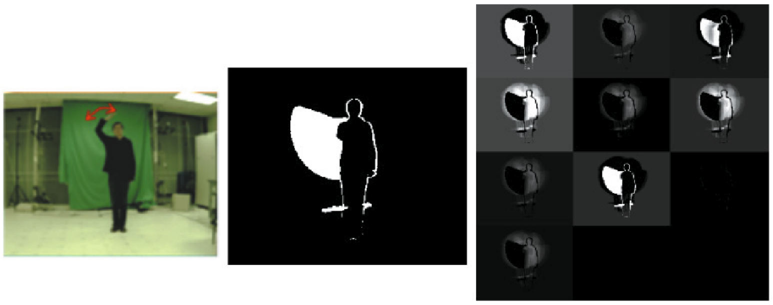


Fig. 16. (Color online) Sample from Keck Gesture dataset: (left) Wave2, (middle) extracted BSTM, (right) activations of the first convolutional layer.

illustrated in the confusion matrix (Fig. 18) and ROC curve (Fig. 19).

5.3. KTH Dataset

To validate and confirm the efficiency of our proposed method, we carried out several investigations using KTH dataset [34]. This is the most used database in the human activity recognition domain consisting of six human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed by 25 subjects in four different scenarios.

The videos in the database include variations in scale, illumination, duration, changes in clothing and changes in viewpoint. The frames are 160 × 120 pixels with a temporal resolution of 25 f/s. Our evaluation protocol consists of using the actions from five subjects for testing, and the remaining subjects for training. Figure 20 shows a sample from KTH dataset, extracted silhouettes and the activations of the convolutional layer of the CNN.

A comparative study against state-of-the-art methods (Table 5) shows that our technique outclasses conventional methods when using KTH dataset and gives

Table 4. Human activity recognition rates obtained in literature and in our approach using Keck Gesture database

Method	Accuracy
Zhuolin Jiang et al. 2012 [33]	97.5% (9 actions)
Our approach	**100% (9 actions)**

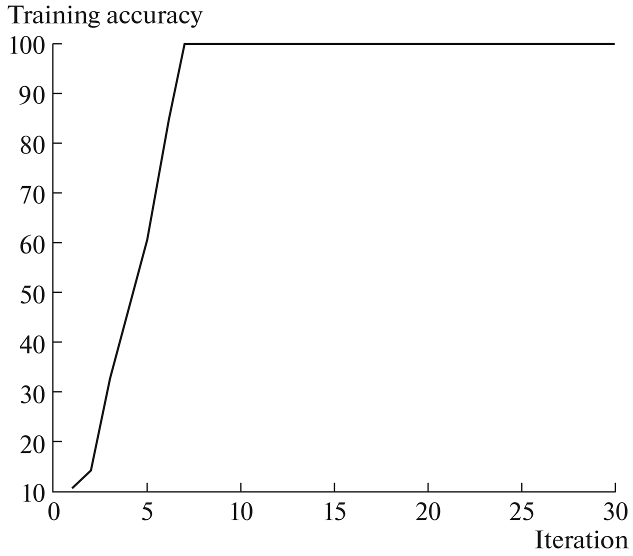


Fig. 17. Training accuracy rate against the current iteration using Keck Gesture database.

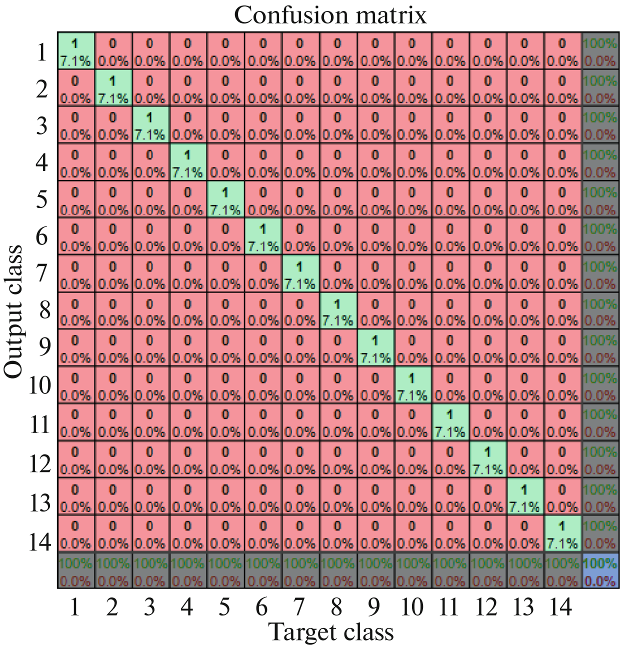
comparable performance against recent deep-learning method.

These results are encouraging because, unlike state-of-the-art methods, our proposed deep-learning architecture is simple. We used only one stream with one convolutional layer, which makes the training process relatively fast (Fig. 21).

Our experimental results show that the quality of the extracted silhouettes has an impact on the global performance of the proposed method, and we believe that efficiency of the proposed method can increase by using more efficient silhouette extraction algorithms.

Table 5. Human activity recognition rates obtained in literature and in our approach using KTH database

Method	Accuracy
Wong and Cipolla. 2007 [41]	86.62%
Niebles et al. 2007 [42]	83.33%
Laptev et al. 2008 [43]	92.10%
Schuldt et al. 2004 [44]	71.70%
Dollar et al. 2005 [3]	81.20%
Bo Chen et al. 2010 [45]	91.13%
Vivek et al. 2015 [46]	93.96%
Lin Sun et al. 2014 [47]	93.10%
Moez B et al. 2015 [23]	94.39%
Baccouche M et al. 2010 [48]	89.40%
Our approach	**92.50%**



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
11	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
12	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	0 0.0%	100% 0.0%
13	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	0 0.0%	100% 0.0%
14	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 7.1%	100% 0.0%
	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%

Fig. 18. (Color online) Confusion matrix using Keck Gesture database.

Confusion matrix in Fig. 22 and ROC curve in Fig. 23 Show that our proposed method gives a 100% accuracy rate for actions: boxing, hand waving and hand clapping, and 95% for walking activity. Most of the classification errors are related to running and jogging activities, this is because of high similarity between the two actions.

6. CONCLUSIONS

In this paper, we presented a simple human activity recognition technique using deep CNN, our method uses human body extracted silhouettes to calculate binary space-time maps (BSTMs) which contain the space-time information of the video stream in a defined time-interval, and CNN to extract features and classification.

Experimental results show that our technique is efficient and produces a perfect recognition rate for both Weizmann and Keck Gesture Dataset. Thus, it outperforms conventional methods when using KTH dataset and provides comparable performance against recent deep—learning methods.

Unlike other viewpoint-based methods, our proposed approach offers the possibility to track and recognize multiple subjects in the same frame because we calculate BSTMs only from the extracted human body, not from the entire frames. The proposed method is simple, efficient, fast to implement and

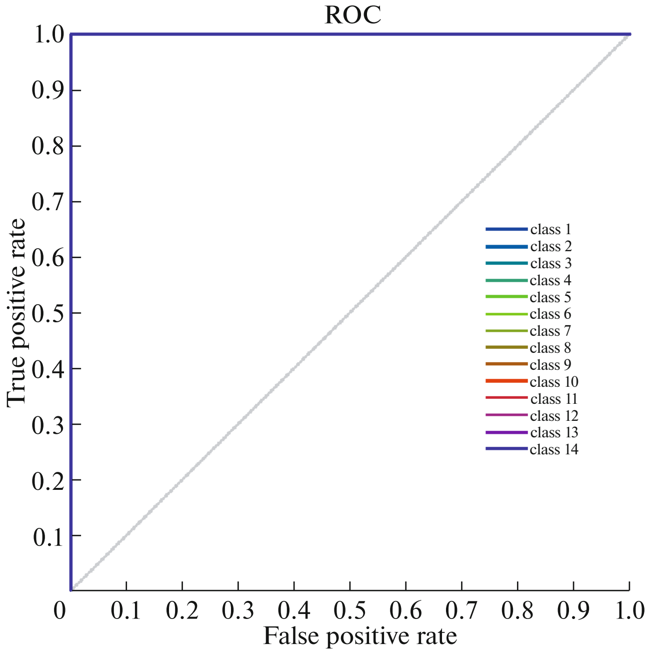


Fig. 19. (Color online) ROC curve using Keck Gesture database.

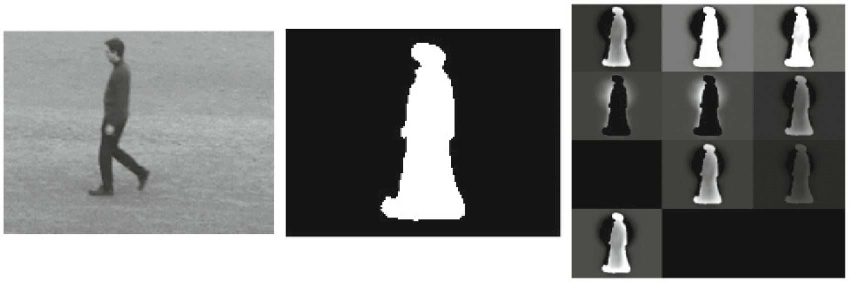


Fig. 20. Sample from KTH dataset: (left) Wave2, (middle) extracted BSTM, (right) activations of the first convolutional layer. ROC curve using.

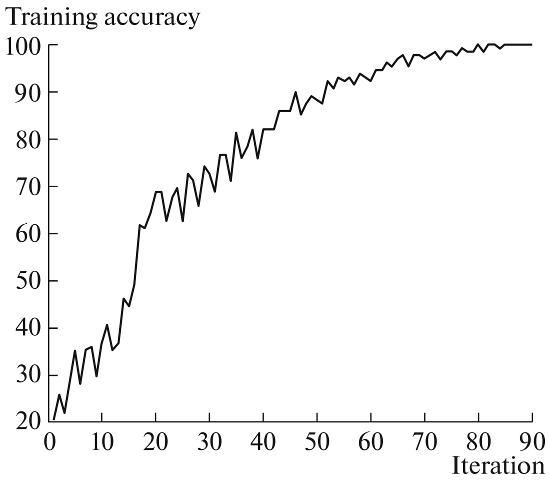
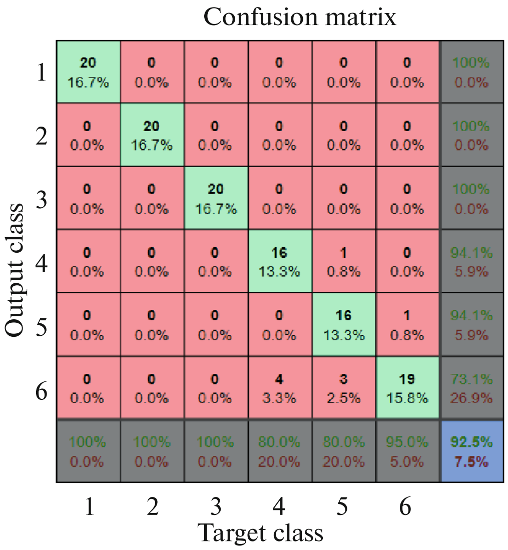


Fig. 21. Training accuracy rate against the current iteration using KTH database.



	1	2	3	4	5	6	
1	20 16.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	20 16.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	20 16.7%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	0 0.0%	0 0.0%	0 0.0%	16 13.3%	1 0.8%	0 0.0%	94.1% 5.9%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	16 13.3%	1 0.8%	94.1% 5.9%
6	0 0.0%	0 0.0%	0 0.0%	4 3.3%	3 2.5%	19 15.8%	73.1% 26.9%
	100% 0.0%	100% 0.0%	100% 0.0%	80.0% 20.0%	80.0% 20.0%	95.0% 5.0%	92.5% 7.5%

Fig. 22. (Color online) Confusion matrix using KTH database.

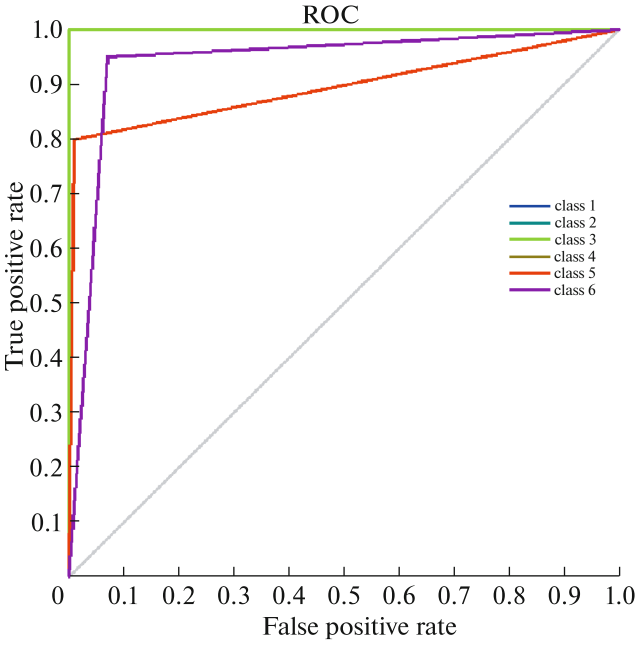


Fig. 23. (Color online) ROC curve using KTH database.

requires less computing time, which makes it suitable for real—time applications.

In future works, we plan to extend our method for multiple viewpoints human activity recognition.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, “A review on video-based human activity recognition,” Comput. **2** (2), 88–131 (2013).
2. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” in *Proc. 10th International Conference on Computer Vision* (*ICCV 2005*) (Beijing, China, 2005), Vol. 1, pp. 1395–1402.
3. P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proc. 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (*VS-PETS*) (Beijing, China, 2005), pp. 65–72.
4. S. Kumari and S. Mitra, “Human action recognition using DFT,” in *Proc. Third IEEE National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics* (*NCVPRIPG 2011*) (Hubli, India, 2011), pp. 239–242.
5. T. Ahmad, J. Rafique, H. Muazzam, and T. Rizvi, “Using discrete cosine transform based features for human action recognition,” J. Image Graphics **3** (2), 96–101 (2015).
6. H. Imtiaz, U. Mahbub, G. Schaefer, et al., “Human action recognition based on spectral domain features,” Procedia Comput. Sci. **60**, 430–437 (2015).
7. D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vision **60** (2), 91–110 (2004).
8. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR 2005*) (San Diego, CA, 2005), Vol. 1, pp. 886–893.
9. W.-L. Lu and J. J. Little, “Simultaneous tracking and action recognition using the PCA-HOG descriptor,” in *Proc. 3rd Canadian Conference on Computer and Robot Vision (CRV 2006)* (Quebec, PQ, Canada, 2006), p. 6.
10. T. Ojala, M. Pjetikäinen, and T. Mäenpää, “Multiresolution grey-scale and rotation invariant texture classification with local binary patterns,” IEEE Trans. Pattern Anal. Mach. Intell. **24** (7), 971–987, 2002.
11. G. Zhao and M. Pietikäinen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” IEEE Trans. Pattern Anal. Mach. Intell. **29** (6), 915–928 (2007).
12. C.-H. Lin, F.-S. Hsu, and W.-Y. Lin, “Recognizing human actions using NWFE-based histogram vectors,” EURASIP J. Adv. Signal Process. **2010**, Article 453064 (2010).
13. A. Nakazawa, H. Kato, and S. Inokuchi, “Human tracking using distributed vision systems,” in *Proc. 14th International Conference on Pattern Recognition* (*ICPR’98*) (Brisbane, Australia, 1998), IEEE, Vol. 1, pp. 593–596.
14. S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, S. Kawato, K. Ebihara, and S. Morishima, “Real-time, 3D estimation of human body postures from trinocular images,” in *Proc. IEEE International Workshop on Modelling People* (*MPeople’99*) (Kerkyra, Greece, 1999), pp. 3–10.
15. F. Huo, E. Hendriks, P. Paclik, and A. H. J. Oomes, “Markerless human motion capture and pose recognition,” in *Proc. 2009 10th Workshop on Image Analysis for Multimedia Interactive Services* (*WIAMIS*) (London, UK, 2009), IEEE, pp. 13–16.
16. S. Sedai, M. Bennamoun, and D. Huynh, “Context-based appearance descriptor for 3D human pose estimation from monocular images,” in *Proc. 2009 Digital Image Computing*: *Techniques and Applications* (*DICTA 2009*) (Melbourne, Australia, 2009), IEEE, pp. 484–491.
17. N. Y. Hammerla, S. Halloran, and T. Ploetz, “Deep, convolutional, and recurrent models for Human Activity Recognition using wearables,” in *Proc. 25th International Joint Conference on Artificial Intelligence* (*IJCAI’16*) (New York, USA, 2016), pp. 1533–1540.
18. C. A. Romao and S.-B. Cho, “Human activity recognition with smartphone sensors using deep learning neural networks,” Expert Syst. Appl. **59**, 235–244 (2016).
19. J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, “Deep convolutional neural networks on multichannel time series for human activity recognition,” in *Proc. 24th International Joint Confer-*

ence on Artificial Intelligence (*IJCAI'15*) (Buenos Aires, Argentina, 2015), pp. 3995–4001.

20. M. A. Alsheikh, A. Selin, D. Niyato, et al., "Deep activity recognition models with triaxial accelerometers," *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence* (*Phoenix, AZ, 2016*): *Artificial Intelligence Applied to Assistive Technologies and Smart Environments*; Technical Report WS-16-01, pp. 8–13.
21. Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signature using deep convolutional neural networks," IEEE Geosci. Remote Sens. Lett. **13** (1), 8–12.
22. T. Dobhal, V. Shitole, G. Thomas, and G. Navada, "Human activity recognition using Binary Motion Image and deep learning," Procedia Comput. Sci. **58**, 178–185 (2015).
23. M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Human Behavior Understanding, Proc. Second International Workshop, HBU 2011*, Ed. by A. A. Salah and B. Lepri, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2011), Vol. 7065, pp. 29–39.
24. P. Wang, W. Li, Z. Gao, J. Zhang, T. Chang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," IEEE Trans. Human-Mach. Syst. **46** (4), 498–509 (2016).
25. Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proc. 23rd ACM International Conference on Multimedia* (*MM'15*) (Brisbane, Australia, 2015), pp. 461–470.
26. P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento, "Exploiting the deep learning paradigm for recognizing human actions," in *Proc. 11th IEEE International Conference on Advanced Video and Signal Based Surveillance* (*AVSS 2014*) (Seoul, South Korea, 2014), pp. 93–98.
27. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.-F. Li, "Large scale video classification with convolutional neural networks," in *Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2014*) (Columbus, OH, 2014), pp. 1725–1732.
28. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," arXiv preprint arXiv:1406.2199 (2014). https://arxiv.org/abs/1406.2199
29. A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Trans. Pattern Anal. Mach. Intell. **23** (3), 257–267 (2001).
30. Y. LeCun and Y. Bengio, "Convolutional networks for images, speech and time-series," in *The Handbook of Brain Theory and Neural Networks*, Ed. by M. A. Arbib (MIT Press, Cambridge, MA, 1995), pp. 255–258.
31. N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst., Man, Cybern. **9** (1), 62–66 (1979).
32. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in Weizmann Database (2007). URL: http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html
33. Z. Jiang, Z. Lin, and L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," IEEE Trans. Pattern Anal. Mach. Intell. **34** (3), 533–547 (2012).
34. C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th International Conference on Pattern Recognition* (*ICPR 2004*) (Cambridge, UK, 2004), IEEE, Vol. 3, pp. 32–36.
35. O. Boiman and M. Irani, "Similarity by Composition," in *Advances in Neural Information Processing Systems 19*: *Proc. Annual Conf. NIPS 2006* (Vancouver, Canada, 2006), pp. 177–184.
36. P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proc. 15th ACM International Conference on Multimedia* (*MM'07*) (Augsburg, Germany, 2007), pp. 357–360.
37. L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *Proc. 2007 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2007*) (Minneapolis, MN, 2007), pp. 1–8.
38. V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Texture based description of movements for activity analysis," in *Proc. Third International Conference on Computer Vision Theory and Applications* (*VISAPP 2008*) (Madeira, Portugal, 2008), Vol. 1, pp. 206–213.
39. V. Kellokumpu, G. Zhao, and M. Pietikäinen, "Human activity recognition using a dynamic texture based method," in *Proc. British Machine Vision Conference* (*BMVC 2008*) (Leeds, UK, 2008), pp. 88-1– 88-10.
40. A. Khelalef, F. Ababsa, and N. Benoudjit, "A simple human activity recognition technique using DCT," in *Advanced Concepts for Intelligent Vision Systems, Proc. 17th International Conference, ACIVS 2016, Lecce, Italy, October 2016*, Ed. by J. Blanc-Talon, C. Distante, W. Philips, et al., Lecture Notes in Computer Science (Springer, Cham, 2016), Vol. 10016, pp. 37–46.
41. S. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *Proc. 2007 IEEE 11th International Conference on Computer Vision* (*ICCV*) (Rio de Janeiro, Brazil, 2007), pp. 1–8.
42. J. C. Niebles, H. Wang, and F.-F. Li, "Unsupervised learning of human action categories using spatial-temporal words," Int J. Comput. Vision **79** (3), 299–318 (2007).
43. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *Proc. 2008 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2008*) (Anchorage, AK, 2008), pp. 1–8.
44. C. Schuldt, I. Laptev, and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," Proc. Int Conf. Pattern Recognition, vol. 3, pp. 32–36, 2004.
45. B. Chen, J.-A. Ting, B. Marlin, and N. de Freitas, "Deep learning of invariant spatio-temporal features from video," in *Deep Learning and Unsupervised Feature Learning Workshop — NIPS 2010* (*24th Annual Confer-*

ence on Neural Information Processing Systems) (Whistler, Canada, 2010), pp. 1–9.

46. V. Vivek, Z. Naifan, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. 2015 IEEE International Conference on Computer Vision* (*ICCV 2015*) (Santiago, Chile, 2015), pp. 4041–4049.

47. L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, and S. Yan, "DL-SFA: Deeply-learned slow feature analysis for action recognition," in *Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2014*) (Columbus, OH, 2014), pp. 2625–2632.

48. M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proceedings of the Second International Workshop on Human Behavior Understanding* (Springer, 2011), pp. 29–39.

Pr. Fakhreddine Ababsa is Full Professor in Augmented Reality and Computer Vision at ENSAM since 2017. He is part of "Institut Image". He received his PhD degree from the University of Evry in 2002. He spent 13 years at IBISC Laboratory (University of Evry) as Associate Professor where he worked in the field of Computer Vision and Robotics. His current research focuses on Computer vision, augmented reality and human computer interactions. He is the author of more than 90 research papers in international academic journals and peer-reviewed conference proceedings. Since 2009, he is awarded a research excellence grant by the French high education ministry. He participated in several national and European research projects. He is member of IEEE, and he has participated in several technical committees of IEEE conferences.

Mr. Khelalef Aziz was born in Jijel, Algeria in 1985. Received his Engineer degree from university of Jijel in 2008, and the M.Sc. degree in signal processing from university of Batna, Algeria, he also obtained the M.Sc. in oil field geophysics from the Algerian institute of petroleum (IAP) in 2014, currently he is a PhD student at university of Batna-2. He worked on image processing, image denoising, image watermarking, seismic data processing, seismic data interpretation, 3D seismic acquisition design; currently he focusses on human activity recognition and deep learning, he participated in several international conference on computer vision field and he is the author of several research papers.

Pr. Nabil Benoudjit was born in Batna, Algeria in 1967. He obtained the State Engineer degree in Electronics in 1991, the M.Sc. degree in Electronics in 1994 from the University of Sétif, Algeria and the PhD degree in Applied Science from the Université catholique de louvain, Belgium in 2003. From 1994 to 1999 and from 2004 to 2010, he has been an Assistant Professor and then Associate Professor of Electronics at the University of Batna, Algeria, where he has taught signal processing, pattern recognition and machine learning. Since 2011, he is a Professor of Electronics at the University of Batna-2, Algeria. He is currently the head of the Machine Learning and Data Mining group of the laboratory (LAAAS), Department of Electronics, University of Batna-2. His present research interests are in the area of machine learning applied to biomedical signals, infrared spectrometers and wind speed (classification and regression). He is co-author of more 35 scientific publications and is a referee of several international journals and conferences.